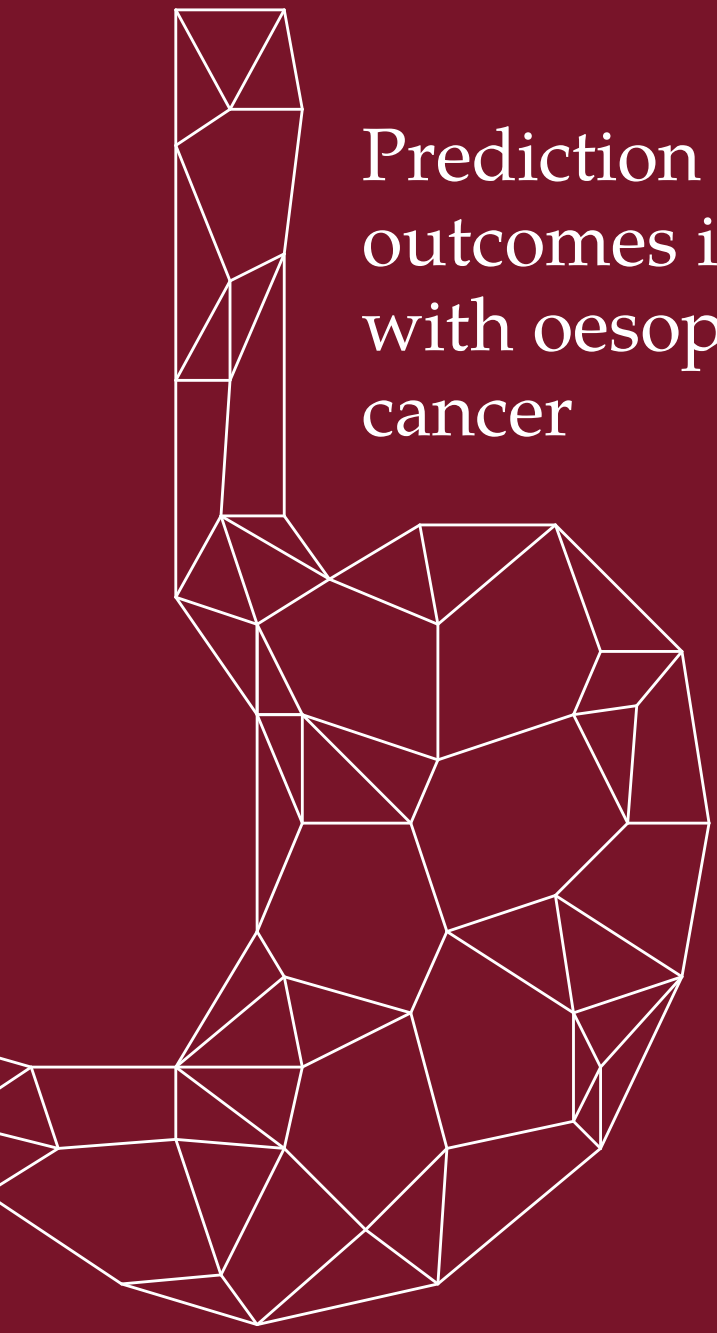
Prediction of treatment outcomes in patients with oesophagogastric cancer

van den Boorn, H.G.

Publication date
2022
Document Version
Final published version



Prediction of treatment outcomes in patients with oesophagogastric cancer

Héctor G. van den Boorn

Prediction of treatment outcomes in patients with oesophagogastric cancer

Héctor G. van den Boorn

Colophon

Cover art	H.G. van den Boorn
Lay-out	H.G. van den Boorn
Printed by	Gildeprint, The Netherlands
ISBN	978-94-6419-375-6

The research for this doctoral thesis received financial assistance from The Dutch Cancer Society (KWF Kankerbestrijding).

Prediction of treatment outcomes in patients with oesophagogastric cancer

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 17 maart 2022, te 13.00 uur

door
Héctor Gabriel van den Boorn

geboren te Curaçao

PROMOTIECOMMISSIE

Promotores	:	prof. dr. H.W.M. van Laarhoven	AMC-UvA
		prof. dr. A.H. Zwinderman	AMC-UvA
Copromotores	:	prof. dr. A. Abu-Hanna	AMC-UvA
		prof. dr. M.A.G. Sprangers	AMC-UvA
Overige leden	:	prof. dr. N.F. de Keizer	AMC-UvA
		prof. dr. M.I. van Berge Henegouwen	AMC-UvA
		prof. dr. P.M.M. Bossuyt	AMC-UvA
		prof. dr. L.V. van de Poll-Franse	Tilburg University
		dr. M. Slingerland	Leids Universitair Medisch Centrum
		dr. M.C. Schut	AMC-UvA

Faculteit der Geneeskunde

Table of Contents

* *These authors contributed equally*

Chapter 1

General introduction

General introduction

Oesophageal and gastric carcinomas occur throughout the world and rank seventh and fifth, respectively, in the number of cancer diagnoses.[1,2] In The Netherlands, every year nearly 4,000 patients are diagnosed with either cancer type.[3] As initial symptoms of oesophagogastric cancer may be mild and hard to timely detect, diagnosis occurs relatively late compared to other types of cancers. As a result, distant spreading of the cancer to other organs (metastases) is present in about 20%-40% of patients at diagnosis.[1,2]

Oesophagogastric cancers are divided in two main groups when it comes to treatment: potentially curable cancers and non-curable metastatic cancers. The potentially curable cancer is characterised by a TNM staging of $T_{1\text{-}4a,X}N_{0\text{-}3}M_0$ where the tumour has not spread to distant organs and has not grown through the wall lining of the oesophagus or stomach. Thanks to these characteristics, surgical removal of the tumour remains a viable option and is considered to be the primary and best treatment. Surgery is often preceded or followed by chemotherapy and/or radiotherapy to prevent recurrence and improve survival. Overall, different combinations of treatment modalities exist for potentially curable cancers, which include among others: surgery alone, (neo)adjuvant treatment with chemoradiotherapy followed or preceded by surgery, definitive chemoradiotherapy and perioperative chemotherapy.[4,5] Although survival in patients with potentially curable cancer is higher than in patients with metastatic disease, outcomes are still relatively poor and the 5-year survival rate does not exceed 50%.[3,6] The potentially curative treatment options provide extended life expectancy, however, potentially at the cost of health-related quality of life (HRQoL) and (life-threatening) toxicity and complications.
Conversely, metastasised cancer cannot be removed by surgery as it has already spread to other organs, and removing the tumour is therefore often not possible. The life expectancy of these patients may span only several months.[3] However, several options are available to prolong life and alleviate symptoms. These include (combinations of) chemotherapy, radiotherapy or placement of a stent.[4,5] Again, these treatments come at a risk, and complications may occur during treatment. Also, HRQoL may be impaired by treatment, and increasing life expectancy and improving HRQoL may not be achieved in all patients. Besides these treatment options targeted at the tumour, best supportive care (BSC) is also an important alternative. BSC does not aim at life-prolongation, but at alleviating symptoms and optimising HRQoL in patients' final stages of life.

In general, it can thus be stated that survival and HRQoL outcomes of cancer treatment are often conflicting. There is not one single treatment that optimises overall survival and HRQoL, while reducing complications at the same time. Given the myriad of treatment options, including BSC, and their potential outcomes, determining the best course of action together with the patient is a difficult and often underestimated task. Weighing the expected gain in quantity vs. health-related quality of life will be influenced by the patient's goals, personal circumstances and expectations, which may differ substantially across patients. For example, some patients want to prolong their life at any cost, while others find HRQoL of utmost importance, whereas yet others want to balance the two. As oesophagogastric cancer patients are relatively old[3], often have comorbidities and have a high incidence of distant metastases, the focus often does not lie solely on life prolongation but also on improving HRQoL, thus complicating treatment choice. Shared decision making with respect to the choice of treatment takes all these factors into consideration.[7]
In shared decision making, patients and oncologists aim to find out what the patient's treatment goals and wishes are and how treatment outcomes (such as survival or HRQoL) are valued by the patient. This can be a difficult task as it may be hard for the patient to think of the best treatment goal in the confusing time following diagnosis, and goals may shift over time. For example, after the initial shock following diagnosis and the prospect of a shortened life, patients may focus more on life prolongation as a treatment goal; however, over time their treatment goals may shift more to improving HRQoL. Therefore, treatment goals need to be addressed repeatedly. Oncologists need to assess the viability of different treatment options by evaluating their potential benefits and risks, and decide, if desired, together with the patient about which treatment to give.
Although this approach may seem obvious, implementation of shared decision making in clinical practice can be complex. One of the challenges is that accurately informing patients about treatment outcomes is a complicated task, as the outcomes may differ from patient to patient and depend on many factors including patient and disease characteristics. Accurate information provision on treatment outcomes, however, is crucial for shared decision making. By providing information on treatment outcomes that is as accurate as possible, subsequent treatments can be chosen that are largely in line with the patient's treatment goals.

Various methods exist to supply accurate information. For example, oncological trial data, based on large groups of patients, can provide treatment outcome information. In these trial reports, outcome information, such as survival or HRQoL, is stratified across different treatment, patient, and tumour characteristics. By combining these individual studies in a meta-analysis, one aims to obtain an accurate as possible estimate of treatment outcomes while also taking potential outlier data into consideration. Although these meta-analyses provide estimates based on large numbers of patients, individual treatment outcomes may differ from the meta-analysed estimates and estimates can be inaccurate for individual patients. Variables such as age, comorbidities and tumour characteristics are often predictive of treatment outcomes such as survival. In prediction models these variables are included as predictors, which in turn could generate more individualised and hence more accurate probability estimates of the outcome. The development of a prediction model, however, requires a sizeable set of individual patient data. The model's predictions provide insight into outcomes, and it can help choose between candidate treatments if these candidates are equally likely to be given. Although these methods aim to supply accurate information, it should be noted that accurate information provision can only be effective in conjunction with the oncologists' expertise, as patients may be underreported in the meta-analyses or prediction models, and not all treatments are viable for all patients.

To summarise, the crucial task of providing accurate information on treatment outcomes is especially difficult due to the many factors on which the outcomes depend, and the often conflicting treatment outcomes (e.g. increased survival at the cost of decreased HRQoL). This complexity impedes accurate information provision and leads to less effective shared decision making in clinical practice. The aim of this thesis is therefore to provide evidence-based information on treatment outcomes, including survival and HRQoL, that is as accurate as possible, by performing meta-analyses and creating prediction models for patients with oesophagogastric cancer. A secondary aim is to present the expected treatment outcomes in a patient-friendly manner to stimulate shared decision making in clinical practice.

Outline of this thesis

In Chapter 2, an overview of existing prediction models for oesophageal and gastric cancer is presented. A systematic review is performed on studies either developing or validating prediction models for oesophagogastric cancer with survival, treatment toxicity and HRQoL as outcomes. This review aims to provide a complete picture on available prediction models for oesophagogastric cancers. The models are evaluated on methodology, bias, applicability and predictive performance. This review also provides recommendations for future models, including those presented in subsequent chapters.

Chapters 3-5 describe the creation and validation of new prediction models for overall survival in patients with oesophagogastric carcinomas, the SOURCE models. In Chapter 3, the novel SOURCE prediction models for overall survival are presented. As there are only few models available for patients with metastatic carcinomas, this study aims to create survival prediction models for patients with metastasised cancers. The models are stratified by tumour location (oesophagus and stomach) and are based on individual patient data from the Netherlands Cancer Registry between 2005-2015. Given patient and tumour characteristics, these models are able to provide survival estimates for a range of treatments. The models' predictive performance and validity are also assessed. Chapter 4 describes the subsequent validation of these models in an external Belgian cohort. Performing this validation on an independent dataset allows to further establish the quality, reliability and generalisability of the models. An update and extension of the SOURCE models is presented in Chapter 5. In this study, the methodology of the earlier models is adapted in order to create more robust models. With the extension of The Netherlands Cancer Registry in 2015, the new models incorporate new data and include important predictors that were unavailable for earlier models. The new models also include survival models for potentially curable cancers. With this addition, the SOURCE models cover the full range of metastatic and potentially curable oesophagogastric cancers. Finally, Chapter 6 describes the creation of the SOURCE-PANC model, a prediction model aimed at predicting overall survival in patients with pancreatic tumours. This chapter serves as a validation of the SOURCE methodology and aims to demonstrate that when following the same steps and considerations in a different setting, a similarly performing prediction model is obtained.

Analyses regarding health-related quality of life in oesophagogastric cancer patients are presented in Chapters 7-9. Chapter 7 describes a systematic review of available literature on HRQoL in metastatic oesophageal and gastric cancer patients. This systematic review is extended by meta-analyses of patient-reported HRQoL scores in the first-line and beyond the first-line setting, stratified by treatment type. In Chapter 8, a similar approach is reported to analyse HRQoL in potentially curable patients stratified by potentially curative treatments such as surgery, definitive chemoradiotherapy and neoadjuvant therapy. In both Chapters 7 and 8, systematic reviews and meta-analyses of the available literature are presented to arrive at evidence-based insight into how HRQoL is influenced by treatment and how it evolves over time. Both analyses aim to provide valuable information to be used in the shared decision making process. With the availability of individual patient data from the POCOP registry[8], a prediction model for HRQoL is developed, as described in Chapter 9. This model is aimed to predict the course of HRQoL after the start of treatment in individual patients. The models are based on a priori information such as patient, treatment and tumour characteristics and are intended to provide more accurate and individualised information regarding HRQoL.

The available analyses and prediction models on treatment outcomes are subsequently combined for use in clinical practice. The creation and subsequent evaluation of a web-interface, aimed at presenting the prediction models and meta-analyses in a patient-friendly manner, are described in Chapter 10. This tool incorporates (a) prediction models for survival that were described in Chapters 3-5, (b) the results from the HRQoL meta-analyses reported in Chapters 7 and 8, and (c) toxicity prevalence for systemic therapy. This web-interface provides options to personalise predictions and visualises the outcomes graphically. The tool is intended for use in clinical practice by the oncologist together with the patient in order to support shared decision making.

This thesis concludes with a general discussion and future perspectives relating to the use and role of prediction models in clinical practice.

References

1. Ferlay, J., Soerjomataram, I., Ervik, M., *et al.* GLOBOCAN 2012 v1. 0, Cancer incidence and mortality worldwide: IARC CancerBase No. 11. 2013. 2014.
2. Bray, F., Ferlay, J., Soerjomataram, I., *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians 68, 394–424 (2018).
3. IKNL. Nederlandse Kankerregistratie (NKR). 2021. `https://iknl.nl/nkr-cijfers`.
4. Lordick, F., Mariette, C., Haustermans, K., *et al.* Oesophageal cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Annals of Oncology 27, v50–v57 (2016).
5. Smyth, E. C., Verheij, M., Allum, W., *et al.* Gastric cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Annals of Oncology 27, v38–v49 (2016).
6. Gavin, A. T., Francisci, S., Foschi, R., *et al.* Oesophageal cancer survival in Europe: A EUROCARE-4 study. Cancer Epidemiology (2012).
7. Stiggelbout, A. M., Van der Weijden, T., De Wit, M. P., *et al.* Shared decision making: Really putting patients at the centre of healthcare. The BMJ (Online) 344 (2012).
8. Coebergh van den Braak, R. R., van Rijssen, L. B., van Kleef, J. J., *et al.* Nationwide comprehensive gastro-intestinal cancer cohorts: the 3P initiative. Acta Oncologica 57, 195–202 (2018).

Chapter 2

Prediction models for patients with oesophageal or gastric cancer

A systematic review and meta-analysis

Héctor G. van den Boorn, Ellen G. Engelhardt, Jessy Joy van Kleef, Mirjam A.G. Sprangers, Martijn G.H. van Oijen, Ameen Abu-Hanna, Aeilko H. Zwinderman, Veerle M.H. Coupé and Hanneke W.M. van Laarhoven

Based on:

H.G. van den Boorn, E.G. Engelhardt, J.J. van Kleef, *et al.* Prediction models for patients with esophageal or gastric cancer: a systematic review and meta-analysis. PLoS ONE 13 (2018).

Abstract

Background

Clinical prediction models are increasingly used to predict outcomes such as survival in cancer patients. The aim of this study was threefold. First, to perform a systematic review to identify available clinical prediction models for patients with oesophageal and/or gastric cancer. Second, to evaluate sources of bias in the included studies. Third, to investigate the predictive performance of the prediction models using meta-analysis.

Methods

MEDLINE, EMBASE, PsycINFO, CINAHL, and The Cochrane Library were searched for publications from the year 2000 onwards. Studies describing models predicting survival, adverse events and/or health-related quality of life (HRQoL) for oesophageal or gastric cancer patients were included. Potential sources of bias were assessed and a meta-analysis, pooled per prediction model, was performed on the discriminative abilities (c-indices).

Results

A total of 61 studies were included (45 development and 16 validation studies), describing 47 prediction models. Most models predicted survival after a curative resection. Nearly 75% of the studies exhibited bias in at least three areas and model calibration was rarely reported. The meta-analysis showed that the averaged c-index of the models is fair (0.75) and ranges from 0.65 to 0.85.

Discussion

Most available prediction models only focus on survival after a curative resection, which is only relevant to a limited patient population. Few models predicted adverse events after resection, and none focused on patient's HRQoL, despite its relevance. Generally, the quality of reporting is poor and external model validation is limited. We conclude that there is a need for prediction models that better meet patients' information needs, and provide information on both the benefits and harms of the various treatment options in terms of survival, adverse events and HRQoL.

Introduction

Worldwide, oesophageal and gastric cancer account for 3.2% and 6.8% of all new cancer cases, respectively. The prognosis is dismal: 1% of patients with oesophageal cancer and 5% of patients with gastric cancer survive at least five years after being diagnosed.[1] However, survival rates for both entities vary greatly[1–4] and metastasis is one of the decisive factors for curative or palliative treatment. In both the curative and palliative setting, patients may choose between various treatment options that differ in terms of efficacy, adverse events and impact on health-related quality of life (HRQoL).

Many patients with potentially curable oesophageal or gastric cancer report loss of HRQoL[5,6] during the first year after surgery, even though patients indicate that an improved HRQoL may be their primary outcome of treatment.[7] Likewise, one in four patients with metastatic oesophageal cancer state that HRQoL is their main treatment goal.[8] Since life prolonging treatment may come at a cost as it may induce adverse events and impair HRQoL[5,6], patients need to be informed at an early stage about the projected survival, adverse events and HRQoL.

To make well-informed treatment choices that match patients' preferences and goals, information about treatment outcomes in terms of survival, treatment-related adverse events and HRQoL is necessary.[9] Statistical prediction models that provide personalised estimates of such outcomes can help inform patients and clinicians consequently supporting shared decision-making. Such statistical models are generally derived from large historical patient cohorts. Examples of such models in oncology are *Adjuvant!*[10] and *PREDICT*[11], which are broadly used in the field of breast cancer. However, a comprehensive overview of available models for oesophageal and gastric cancer, and their predictive performance is currently lacking. Therefore, the aim of this systematic review was first to provide an overview of published prediction models that provide personalised estimates of survival probabilities (i.e., overall, disease-specific, progression-free or disease-free survival), the probability of developing treatment-related adverse events, and/or the impact of treatment on HRQoL. Secondly, we aimed to examine the quality of the development and validation studies conducted for the identified prediction models. Finally, we evaluated the reported performance of the prediction models in terms of discriminative ability and calibration.

Methods

Systematic literature search

A systematic literature search was performed to identify all relevant publications in the bibliographic databases MEDLINE, EMBASE, PsycINFO, CINAHL, and The Cochrane Library. To increase the relevance of the findings of this review for current clinical practice, we only included papers published from January, 1st 2000 up to February 6th, 2017. Search terms for 'oesophageal cancer' or 'gastric cancer' were used in combination with search terms for 'prediction model', 'survival', 'adverse events' and 'health-related quality of life' (see Supplementary Table 1 for the detailed search strategy). The reference lists of relevant identified articles were also searched for additional relevant publications.

The aim of our search was to identify prediction models that provide personalised estimates of survival, the probability of experiencing an adverse event and/or the impact of disease or treatment on HRQoL for oesophageal and gastric cancer patients. Models intended to support treatment decisions in both the curative or the palliative setting were eligible for inclusion. Studies validating models in patients with oesophageal or gastric cancer that were not originally developed for use in these populations, were also eligible for inclusion. Also, only papers published in English were assessed. We excluded studies describing prediction models that aimed to classify patients into risk categories (such as 'low risk' and 'high risk'), rather than providing personalised estimates of outcome probability. Although risk categories may be useful for discriminating between outcome severity, it is difficult to quantify the calibration of such prediction models (i.e., how does the expected outcome compare to the actual observed outcome). This is an important aspect of model validation, as the absolute outcome probabilities are needed to determine model fit, and therefore, the quality of the model.

The selection process consisted of two phases. First, all titles and abstracts were screened by two reviewers (HvdB and EE) independently. Discrepancies were resolved through consensus, and when necessary by consulting a third arbiter (HvL). Studies were also selected if eligibility could not be determined on the basis of the titles and abstracts. In the second phase, two reviewers (HvdB and EE) independently screened full texts of the studies selected in phase one to determine eligibility conclusively.

Data extraction

Data were extracted from the full text papers according to the CHARMS[12] statement, which provides a data extraction checklist for systematic reviews of prediction models. Extracted data included information about the type of article, study design, data source, characteristics of the population, aim of the model, type of outcome, sample size, methods used and presentation of the final prediction model. Model performance was also extracted and categorised as development performance (obtained when using the development dataset), internal validation performance (obtained when using data from a population similar to that of the development set), and external validation performance (when the data used differs temporally, geographically etc. from the development set). Model performance was described using measures for discriminative ability and measures for calibration. Discriminative ability is defined as a model's ability to differentiate between patients who experience an event (such as death or an adverse event) and those who do not.[13] This can be quantified by calculating an index of predictive discrimination, the concordance index (c-index). This c-index typically has values ranging from 0.5 (no discrimination at all) to 1 (perfect discrimination), and is the generalisation of the area under the curve, a well-known measure of discrimination. Typically c-indices can be interpreted by the following rule of thumb: 0.5–0.6 no discrimination, 0.6–0.7 poor, 0.7–0.8 fair, 0.8–0.9 good and 0.9–1 excellent discrimination. Model calibration, in contrast, conveys the goodness of fit, i.e., the congruence between observed and average predicted outcomes.[13] Calibration can be displayed visually in a calibration plot.

The levels of evidence of the discriminatory accuracy of the prediction model as described by Reilly and Evans[14], indicates how extensively a prediction model has been validated and to what extent a model is ready for clinical use. Level 1 refers to model development, level 2 to narrow validation, level 3 to broader validation and level 4 and 5 to respectively narrow or broad impact analysis. Each identified study was categorised according to the Reilly-Evans levels. For the assessment of bias, there are no established checklists specifically designed for use in prediction modelling studies. We therefore created a classification system for several areas of possible bias, which were derived from the TRIPOD-statement (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis).[15] Supplementary Table 2 presents an overview of the classification system used for potential risk of bias.

Data extraction was performed by two researchers (HvdB and EE). First, a subset of 10 articles was used as a training set. The training set was coded by both researchers independently and discrepancies in coding were resolved during a consensus meeting. The percentage overall agreement between the two coders was approximately 90% across individual items. The coding scheme was revised where necessary as a result of the training set findings. Thereafter, each researcher coded half of the remaining articles. Classification of the potential for bias was done in two stages; each researcher made notes of potential sources of bias per category separately, and together they (HvdB and EE) then categorised the identified potential sources of bias. The bias was determined in six areas: population-related (such as selection bias), predictor-related (such as ill-defined predictors), outcome-related (such as an unclear outcome), sample size-related, missing data-related (such as only complete case analysis) and statistical analysis-related (such as underreporting of statistics).

Bias analyses

Descriptive analyses were used to summarise study and model characteristics. We expected that the higher the impact factor of a journal in which the study was published, the more stringent the internal screening and peer review procedures would be and, hence, the lower the risk of bias. Further, we hypothesised that the higher the impact factor of the journal a prediction model was published in, the better its performance in terms of c-index would be. Both hypotheses were assessed through the Spearman rank correlation between the journal impact factor[16] (in the year of publication, or the closest to publication year available) and the reported c-index as well as between journal impact factor and the potential sources of bias (assessed using the classification of potential sources of bias presented in Supplementary Table 2), respectively. Due to differences in oesophageal carcinoma histology in different geographical populations[17], we examined whether models were constructed and validated with patient cohorts from different continents using the Fisher's exact test. Finally, we hypothesised that the reported c-indices would be larger during model development than during validation due to overfitting. This was assessed using a one-tailed Wilcoxon signed-rank test. These analyses were performed in the R-studio environment with R version 3.3.3 (R Foundation for Statistical Computing, Vienna, Austria, https://www.r-project.org).

Meta-analyses of c-indices

To gain insight in the discriminative abilities of the prediction models, we performed meta-analyses. The c-indices were pooled per prediction model using random effects modelling for models for which at least two concordance indices were available. Analyses were performed using linear restricted maximum-likelihood estimation. In most articles, the c-index confidence interval or variance was not reported. In those cases, the study weights in the meta-analysis were determined as the inverse square root of the sample size. The logistic transformation as described in Kottas et al.[18] was applied to all c-index estimates during calculations and then transformed back; this procedure ensures that all estimates are bounded by 0 and 1 after pooling, which is a property of the c-index. These analyses were performed using the *metafor* package in the R-studio environment (R version 3.3.3).

Results

A total of 8,963 articles was identified, of which 61 were eligible for inclusion in this systematic review (Figure 1). These studies described a grand total of 47 prediction models for patients with oesophageal or gastric cancer. Two studies describing the development of a prediction model, were not included in our systematic review due to the publication year (POSSUM[19]), and incorrect patient population (P-POSSUM[20]). The remaining 45 development studies are shown in Table 1. Further, we found 16 validation studies on a total of 10 prediction models. These studies are shown in Table 2.

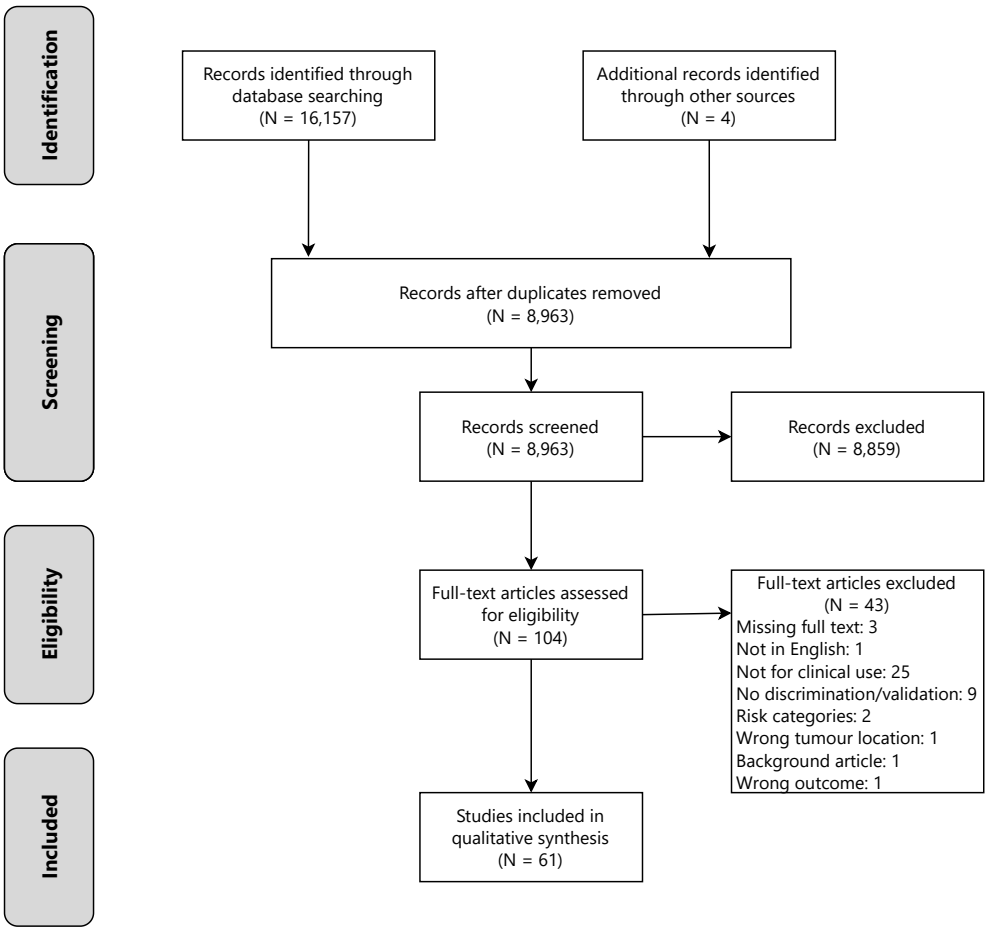


Figure 1: Overview of study selection according to the 'Preferred Reporting Items for Systematic reviews and Meta-Analyses' (PRISMA) statement[21].

Study	N	Country	Tumour location	Treatment intention	Outcome	Model c-indices	Model presentation	Reilly-Evans level
Biglarian 2011[22]	300	Iran	Stomach	Unclear	OS	OS: 0.88 (dev), 0.92 (int)	None	1
Cao 2016[23]	4,281	USA, China	Oesophagus	Unclear	CSS	DSS: 0.72 (dev), 0.699 (ext)	Nomogram	2
Chen, S. 2016[24]	308	China	Oesophagus	Curative	DSS	DSS: 0.688 (dev)	Nomogram	1
Deans 2007[25]	220	UK	Oesophagogastric	Curative/Palliative	OS	OS: 0.84 (dev), 0.85 (dev)	Formula	1
Dhir 2012[26]	14,235	USA	Stomach	Curative/Palliative	POM	POM: 0.75 (ext)	Nomogram	2
Dikken 2013[27]	1,642	USA/NL	Stomach	Curative	DSS	DSS: 0.77 (dev)	Nomogram	1
Duan 2016[28]	328	China	Oesophagus	Curative	OS, DFS	OS: 0.71 (dev), 0.77 (int); DFS: 0.71 (dev), 0.65 (int)	Nomogram	1
Eil 2014[29]	824	USA	Oesophagus	Unclear	OS	OS: 0.72 (dev)	Online tool	1
Eom 2015[30]	1,579	Korea	Stomach	Curative	OS	OS: 0.831 (ext)	Nomogram	3
Filip 2015[31]	167	Italy	Oesophagus	Unclear	AE	AE: 0.8 (dev)	Formula	3
Fischer 2016[32]	4,882	UK	Oesophagogastric	Curative	POM, AE	POM: 0.698 (dev), 0.694 (dev); AE: 0.631 (dev)	Formula	1
Fuccio 2016[33]	267	Italy	Oesophagus	Curative/Palliative	AE	AE: 0.617 (dev), 0.617 (dev), 0.622 (dev)	Table	1
Gabriel 2017[34]	7,179	USA	Oesophagus	Curative	OS	OS: 0.656 (dev), 0.669 (dev), 0.63 (int), 0.682 (int)	Formula	1
Haga 2015[35]	762	Japan	Stomach	Unclear	OS	OS: 0.89 (dev)	Formula	1
Han 2012[36]	5,300	Korea, Japan	Stomach	Unclear	OS	OS: 0.78 (int), 0.79 (ext)	Nomogram	2
Hirabayashi[37]	3,085	Japan	Stomach	Curative	OS	OS: 0.68 (ext)	Nomogram	2
Jiang 2016[38]	125	China	Stomach	Unclear	OS	OS: 0.868 (int), 0.698 (int), 0.84 (int), 0.786 (int), 0.836 (ext), 0.669 (ext), 0.832 (ext), 0.749 (ext)	Nomogram	2
Jung 2013[39]	239	Korea	Oesophagus	Palliative	OS	OS: 0.69 (dev)	Nomogram	1
Kattan 2003 (MSKCC)[40]	1,039	USA	Stomach	Curative	DSS	DSS: 0.8 (dev)	Nomogram, online tool	3
Kim, Y. 2015[41]	719	USA	Stomach	Curative	OS, DFS	OS: 0.711 (dev), 0.691 (ext); DFS: 0.702 (dev), 0.685 (ext)	Nomogram	1
Kunisaki 2016[42]	52,770	Japan	Stomach	Unclear	AE	AE: 0.797 (int), 0.784 (int), 0.748 (int), 0.832 (int), 0.728 (int), 0.7 (int), 0.779 (int), 0.658 (int)	Formula	2
Kurita 2015[43]	33,917	Japan	Stomach	Curative	POM	POM: 0.785 (dev), 0.798 (int)	None	1
Lagarde 2007b[44]	364	Unclear	Oesophagus	Curative	DSS	DSS: 0.77 (dev)	Nomogram	2
Lagarde 2008a[45]	663	Netherlands	Oesophagus	Curative	AE	AE: 0.65 (dev), 0.666 (int)	Nomogram	3
Lai 2009[46]	2,923	Korea	Stomach	Curative	DFS	DFS: 0.79 (dev)	None	2
Liu, J. 2016a[47]	817	China	Stomach	Unclear	OS	OS: 0.79 (ext)	Nomogram	1
Liu, J. 2016b[48]	2,770	USA, China	Stomach	Curative	DSS	DSS: 0.73 (int), 0.76 (ext)	Nomogram	2
Liu, J.S. 2015[49]	326	China	Oesophagus	Curative	DSS	DSS: 0.72 (dev)	Nomogram	1
Marrelli 2005[50]	536	Italy	Stomach	Curative	DFS	DFS: NA (dev)	Formula	2
Mohammadzadeh 2015[51]	194	Iran	Stomach	Unclear	OS	OS: 0.8 (dev), 0.79 (int)	Decision tree	1
Muneoka 2016[52]	207	Japan	Stomach	Curative	DFS	DFS: 0.8 (dev)	Nomogram, online tool	1

Table 1 – *continued from previous page*

Study	N	Country	Tumour location	Treatment intention	Outcome	Model c-indices	Model presentation	Reilly-Evans level
Shao 2015[53]	633	China	Oesophagus	Curative	OS	OS: 0.77 (dev), 0.77 (dev), 0.76 (int), 0.77 (int)	Nomogram	1
Shapiro 2016[54]	626	Netherlands	Oesophagus	Curative	OS	OS: 0.63 (dev)	Nomogram	1
Shiozaki 2016[55]	64	USA	Oesophagogastric	Palliative	OS	OS: 0.61 (dev)	Nomogram	1
Song 2014[56]	805	Korea	Stomach	Curative	DSS	DSS: 0.87 (dev), 0.84 (int)	Nomogram, formula	1
Steyerberg 2006[57]	1,327	USA, Netherlands	Oesophagus	Unclear	POM	POM: 0.66 (dev), 0.7 (ext), 0.56 (ext), 0.66 (ext)	Formula	3
Su 2015[58]	797	China	Oesophagus	Unclear	OS	OS: 0.73 (dev), 0.715 (int)	Nomogram	1
Suzuki 2012[59]	196	USA	Oesophagus	Unclear	OS, DFS	OS: 0.7 (dev); DFS: 0.77 (dev)	Nomogram	1
Tekkis 2004 (O-POSSUM)[60]	1,042	UK	Oesophagogastric	Curative/Palliative	POM	POM: 0.8 (dev)	Formula	3
Tu 2017[61]	3,632	China	Stomach	Curative	AE	AE: 0.68 (dev)	Nomogram	1
Woo 2016[62]	11,851	Korea, Japan, China	Stomach	Curative/Palliative	OS	OS: 0.824 (dev), 0.842 (ext), 0.868 (ext), 0.839 (ext), 0.798 (ext)	Formula	3
Yang 2013[63]	319	China	Oesophagus	Curative	REC	Not available	Formula	1
Yu 2016[64]	1,004	China	Oesophagus	Curative	OS	OS: 0.7 (dev)	Nomogram	1
Zhao 2016[65]	510	China	Stomach	Curative	OS	OS: 0.834 (dev), 0.809 (int)	Nomogram	1
Zhou, Z. 2015[66]	953	USA, China	Oesophagus	Curative	OS	OS: 0.69 (dev), 0.75 (ext)	Nomogram	2

Table 1: Overview of selected studies which describe the creation of a novel prediction model.
The type of validation is indicated in brackets with the reported c-index; dev: development c-index, int: internal validation, ext: external validation; N: sample size used for training; DSS: disease-specific survival; POM: post-operative mortality; OS: overall survival; AE: adverse events; DFS: disease-free survival; REC: cancer recurrence.

Study	Validation of	N	Country	tumour location	Treatment intention	Outcome	Model c-indices	Reilly-Evans level
Ashfaq 2015[67]	MSKCC[40]	6,954	USA	Stomach	Curative	DSS	0.68	3
Bosch 2011[68]	P-POSSUM[20], O-POSSUM[60]	278	Netherlands	Oesophagus	Curative	POM	0.766, 0.756	3
Chen, D. 2013[69]	MSKCC[40]	979	China	Stomach	Curative	DSS	0.74	3
D'Journo 2016[70]	Steyerberg 2006[57]	1,039	France	Oesophagus	Unclear	OS	0.63, 0.64, 0.63	3
Dikken 2014[71]	MSKCC[40]	139	USA	Stomach	Unclear	DSS	0.64	3
Grotenhuis 2010[72]	Lagarde 2008a[45]	777	Netherlands	Oesophagus	Curative	AE	0.64	3
Kim, J.H. 2012[73]	Lai 2009[46]	930	Korea	Stomach	Curative	DFS	0.7	2
Lagarde 2007a[74]	O-POSSUM[60]	663	Netherlands	Oesophagus	Curative	POM	0.6	3
Lagarde 2008b[75]	Lagarde 2007b[44]	382	Belgium	Oesophagus	Curative	DSS	0.76	2
Marrelli 2015[76]	Marrelli 2005[50]	635	Italy	Stomach	Curative	REC	0.889	2
Nagabhushan 2007[77]	P-POSSUM[20], O-POSSUM[60]	313	UK	Oesophago-gastric	Curative	POM	0.68, 0.61	3
Novotny 2006[78]	MSKCC[40]	862	Germany	Stomach	Curative	DSS	0.77	3
Peeters 2005[79]	MSKCC[40]	459	Netherlands	Stomach	Curative	DSS	0.77	3
Reim 2015[80]	Eom 2015[30]	908	Germany	Stomach	Curative	OS	0.761	3
Zafirellis 2002[81]	POSSUM[19]	204	UK	Oesophagus	Curative-/Palliative	OS, AE	OS: 0.62 AE: 0.55	3
Zhou, M.L. 2016[82]	MSKCC[40]	150	China	Stomach	Curative	DSS	0.657	3

Table 2: Overview of studies which externally validate prediction models.
N: sample size used for validation; DSS: disease-specific survival; POM: post-operative mortality; OS: overall survival; AE: adverse events; DFS: disease-free survival; REC: cancer recurrence.

Of the models described in the 45 development studies, six predict adverse events; one predicts the recurrence of malignancy; and most studies (N = 39) predict various types of survival (six disease-free survival, eight disease-specific survival, 23 overall survival and five post-operative mortality). None of the studies predict HRQoL and none predict more than one outcome, i.e., no model predicts both the harms and benefits of the treatments of interest. The majority of studies (N = 28) used a nomogram to present the prediction model, while others (N = 13) used a formula as a presentation method (see Table 1). Three prediction models were also available online. A graphical overview of the outcomes per prediction model is given in Figure 2, and includes depictions of each model's Reilly-Evans level of evidence and discriminatory accuracy.

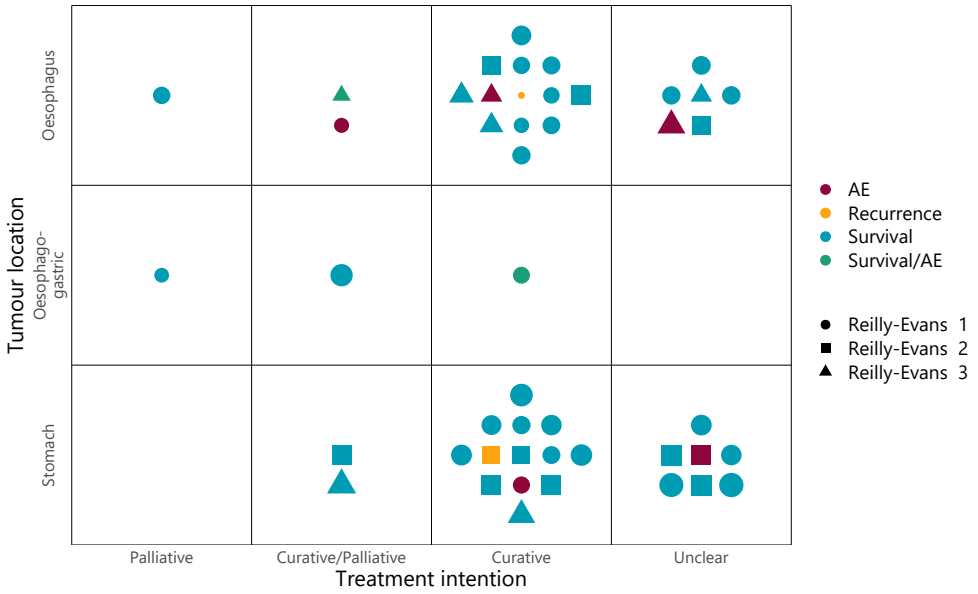


Figure 2: Overview of included prediction models.
The shape indicates the type of study and the size of shapes indicate the pooled c-index. Larger sizes of shapes indicate higher c-indices. AE: adverse event; Reilly-Evans: levels of evidence on the discriminatory accuracy of the prediction model described by Reilly and Evans[14], which indicate how extensively a prediction model has been validated and to what extent a model is ready for clinical use.

Table 3 provides an overview of the selected studies. Most models underwent only limited validation, as the majority of development models were not validated further in later studies. This is expressed by the Reilly and Evans levels of evidence.[14] In 84% of the development studies the two lowest Reilly and Evans levels, namely 1 or 2, were scored indicating only narrow validation. The validation studies are limited to a select group of prediction models, which are validated more extensively. These are the prediction models developed by Eom 2015[30], Lagarde 2007[44], Lagarde 2008[45], Lai 2009[46], Marelli 2005[50], Steyerberg 2006[57], the MSKCC[40], and the Possum[19], O-Possum[60], and P-Possum[20] models. This more extensive validation resulted in a majority of these models having a Reilly and Evans level of 3.

Table 3 also indicates the study patient distribution across the continents. This differs significantly between development and validation studies ($p = 0.003$), indicating that different populations are used for model development and for validation. This difference is especially pronounced between Asia and Europe ($p < 0.001$). Models were more often developed in Asian than in European populations (56.8% vs. 18.2% respectively), however, fewer validation studies were conducted in Asian than in European populations (18.8% vs 68.8% respectively). The development and validation studies mostly concerned prediction outcomes before or after resection (89% and 100% respectively), and were mostly aimed at patients treated with curative intent (56% and 81.2% respectively).

	Development studies	**Validation studies**	**p-value**
N	45	16	
Reilly-Evans level (%)			
1	27 (60.0)	0 (0.0)	
2	11 (24.4)	3 (18.8)	
3	7 (15.6)	13 (81.2)	
Continent of patient population (%)			0.003
Asia	25 (56.8)	3 (18.8)	
Europe	8 (18.2)	11 (68.8)	
North-America	10 (22.7)	2 (12.5)	
North-America and Europe	1 (2.3)	0 (0.0)	
Intended time of model use (%)			0.857
After adjuvant chemotherapy	1 (2.2)	0 (0.0)	
After consolidation therapy	1 (2.2)	0 (0.0)	
After definitive chemoradiation	1 (2.2)	0 (0.0)	
After resection	32 (71.1)	14 (87.5)	
At diagnosis	1 (2.2)	0 (0.0)	
Before definitive chemotherapy	1 (2.2)	0 (0.0)	
Before resection	5 (11.1)	2 (12.5)	
Before/after resection	3 (6.7)	0 (0.0)	
Curative/palliative setting (%)			0.316
Curative	25 (55.6)	13 (81.2)	
Curative/palliative	5 (11.1)	1 (6.2)	
Palliative	2 (4.4)	0 (0.0)	
Unclear	13 (28.9)	2 (12.5)	
Calibration method (%)			0.045
Calibration plot	23 (51.1)	6 (37.5)	
Statistical analysis	2 (4.4)	4 (25.0)	
Calibration plot and statistical analysis	6 (13.3)	4 (25.0)	
None	14 (31.1)	2 (12.5)	

Table 3: Overview of study characteristics in development and validation studies.

Bias analyses

We analysed several areas of possible bias of the studies, which are shown in Table 4. The exact definitions of the biases are presented in Supplementary Table 2. Of all selected studies, population-related bias occurred in 61%, predictor-related bias in 43%, outcome-related bias in 43%, sample size-related bias in 38%, missing data-related bias in 89% and statistical analysis-related bias in 66%. All studies have a bias in at least one area. Due to poor or inconsistent reporting, it was difficult to extract pertinent study information. For example, treatment intent was not reported in most articles. In such cases intent was deduced from other available information such as the presence of metastatic disease. However, in 15 studies the treatment intent could not be established. Also, unclear descriptions of treatment and patient characteristics limited our ability to evaluate the risk of bias. The potential source of bias that was most difficult to evaluate due to poor reporting, concerns the handling of missing data. Although few studies report that their dataset was complete, most studies did not mention whether this was the case and how they handled missing data (e.g., via multiple imputation). Further, in many studies, it was unclear what outcome was being predicted. For example, authors mention 'survival' as an outcome[51], but it remained unclear whether overall survival or disease-specific survival was implied.

In most studies the model calibration was poorly reported. Although 45 out of 61 studies described some form of calibration, only 16 studies performed a formal statistical calibration analysis to support whether the predicted risk matched the observed risk. None of the studies determined the calibration slope and intercept (which represents the systematic over- or underprediction of risk).

Finally, we also investigated whether the impact factor of the journal in which the study was published influenced the amount of bias. We found no significant correlation between journal impact factor and the risk of population-related bias ($\rho = 0.09$, p = 0.51), predictor-related bias (ρ = -0.12, p = 0.37), outcome-related bias ($\rho = 0.17$, p = 0.20), sample size-related bias ($\rho = 0.13$, p = 0.32), missing data-related bias ($\rho = 0.03$, p = 0.79) or statistical analysis-related bias ($\rho = 0.03$, p = 0.80). When we assessed whether models published in high impact journals performed better in terms of discriminative ability, again, we found no relation between the impact factor of the journal and the reported c-index ($\rho = 0.15$, p = 0.11).

Study	Subject bias	Predictor bias	Outcome bias	Sample size bias	Missing data bias	Statistical analysis bias
Ashfaq 2015[67]	–	+	+	++	– –	–
Biglarian 2011[22]	–	–	?	–	– –	– –
Bosch 2011[68]	+	+	–	–	–	–
Cao 2016[23]	–	+	?	+	–	+
Chen, D. 2013[69]	+	+	+	+	– –	–
Chen, S. 2016[24]	– –	+	+	–	–	–
D'Journo 2016[70]	– –	+	+	+	–	+
Deans 2007[25]	+	–	+	–	–	–
Dhir 2012[26]	–	–	?	++	–	+
Dikken 2013[27]	–	–	+	+	– –	–
Dikken 2014[71]	–	+	–	–	– –	–
Duan 2016[28]	+	+	–	+	– –	+
Eil 2014[29]	– –	–	– –	+	– –	–
Eom 2015[30]	–	–	+	–	– –	+
Filip 2015[31]	– –	–	– –	–	–	–
Fischer 2016[32]	+	–	+	+	+	–
Fuccio 2016[33]	–	+	– –	–	– –	–
Gabriel 2017[34]	–	–	+	++	– –	–
Grotenhuis 2010[72]	+	–	–	+	+	–
Haga 2015[35]	–	+	– –	+	–	–
Han 2012[36]	–	+	?	+	– –	+
Hirabayashi 2014[37]	+	+	?	+	– –	–
Jiang 2016[38]	– –	+	– –	–	–	–
Jung 2013[39]	–	+	–	–	–	+
Kattan 2003[40]	–	–	+	+	– –	–
Kim, J.H. 2012[73]	+	–	–	+	– –	–
Kim, Y. 2015[41]	+	+	– –	+	–	–
Kunisaki 2016[42]	–	+	–	++	–	+
Kurita 2015[43]	+	–	–	+	–	–
Lagarde 2007a[74]	+	+	–	+	–	–
Lagarde 2007b[44]	– –	+	+	+	?	–
Lagarde 2008a[45]	+	+	–	+	+	–
Lagarde 2008b[75]	+	–	+	–	+	–
Lai 2009[46]	+	–	+	+	– –	–
Liu, J. 2016a[47]	–	+	?	+	–	+
Liu, J. 2016b[48]	+	–	?	+	– –	+
Liu, J.S. 2015[49]	– –	–	–	–	–	–
Marrelli 2005[50]	+	–	?	–	–	–
Marrelli 2015[76]	+	+	+	+	– –	+
Mohammadzadeh 2015[51]	– –	–	– –	–	– –	– –
Muneoka 2016[52]	–	+	–	–	–	–
Nagabhushan 2007[77]	+	+	–	–	– –	–
Novotny 2006[78]	–	+	+	+	– –	–
Peeters 2005[79]	–	–	+	–	– –	–
Reim 2015[80]	– –	+	+	+	+	+
Shao 2015[53]	+	+	+	+	–	+
Shapiro 2016[54]	+	–	–	+	–	–
Shiozaki 2016[55]	+	+	+	– –	– –	–
Song 2014[56]	–	–	– –	+	– –	+
Steyerberg 2006[57]	–	–	+	+	?	–
Su 2015[58]	– –	+	+	+	– –	+
Suzuki 2012[59]	– –	+	?	–	–	–

Table 4 – *continued from previous page*

Study	Subject bias	Predictor bias	Outcome bias	Sample size bias	Missing data bias	Statistical analysis bias
Tekkis 2004[60]	+	−	+	+	−	−
Tu 2017[61]	+	+	− −	+	− −	−
Woo 2016[62]	−	+	?	++	−	+
Yang 2013[63]	−	+	+	−	−	−
Yu 2016[64]	+	+	+	+	−	+
Zafirellis 2002[81]	−	+	− −	−	− −	+
Zhao 2016[65]	−	−	− −	−	−	+
Zhou, M.L. 2016[82]	−	+	+	−	− −	+
Zhou, Z. 2015[66]	−	−	−	+	−	+

Table 4: Overview of areas of bias in the included studies.
A minus sign indicates possible areas of bias; a question mark indicates that bias could not be determine; a plus sign indicates a lack of bias.

Meta-analyses of c-indices

Results of the meta-analysis of available c-indices of corresponding prediction models are shown in Figures 3A and 3B. Results are pooled per prediction model and are indicated by diamonds. Overall, the meta-analysis highlights that there is great uncertainty about the predictive performances of available models, given the large confidence intervals (with ranges >0.1) in most pooled estimates. Furthermore, the pooled estimates show that the models vary in discriminating ability, ranging from 0.65 (poor discrimination) to 0.85 (good discrimination), with an average pooled estimate of 0.75 (fair discrimination).

To investigate whether model overfitting occurs, that is the discriminative ability of a model is overestimated during training, we examined the difference in model c-indices. It was found that the discriminative ability of the model was indeed larger ($p = 0.01$) in development (average c-index: 0.76) than in validation studies (average c-index = 0.73).

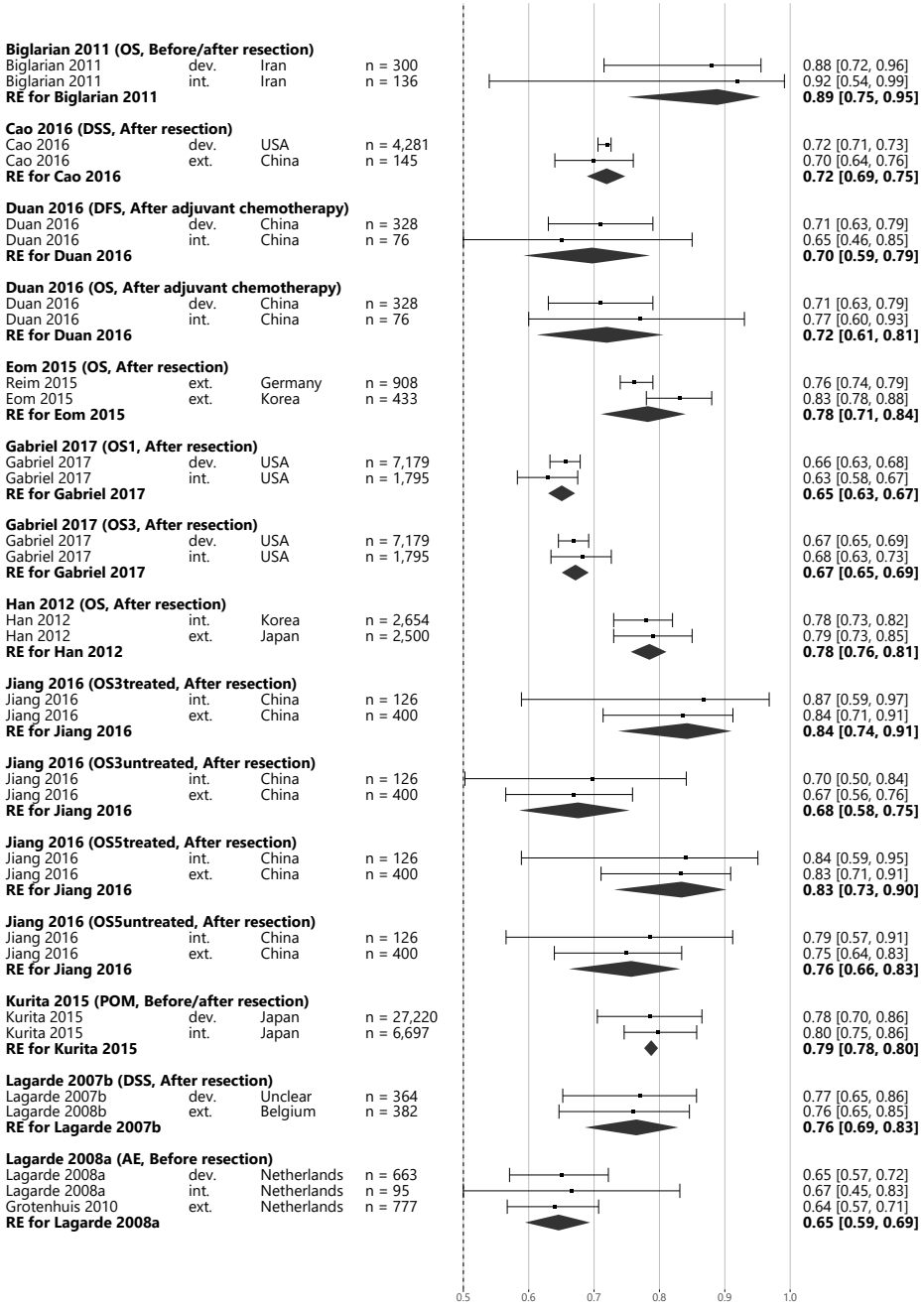


Figure 3A: Random effects meta-analyses of the discriminative abilities (c-indices) of the identified prediction models.

DSS: disease-specific survival; POM: post-operative mortality; OS: overall survival; AE: adverse events; DFS: disease-free survival; REC: cancer recurrence; dev: development c-index, int: internal validation; ext: external validation; RE: random effects meta-analysis estimate.

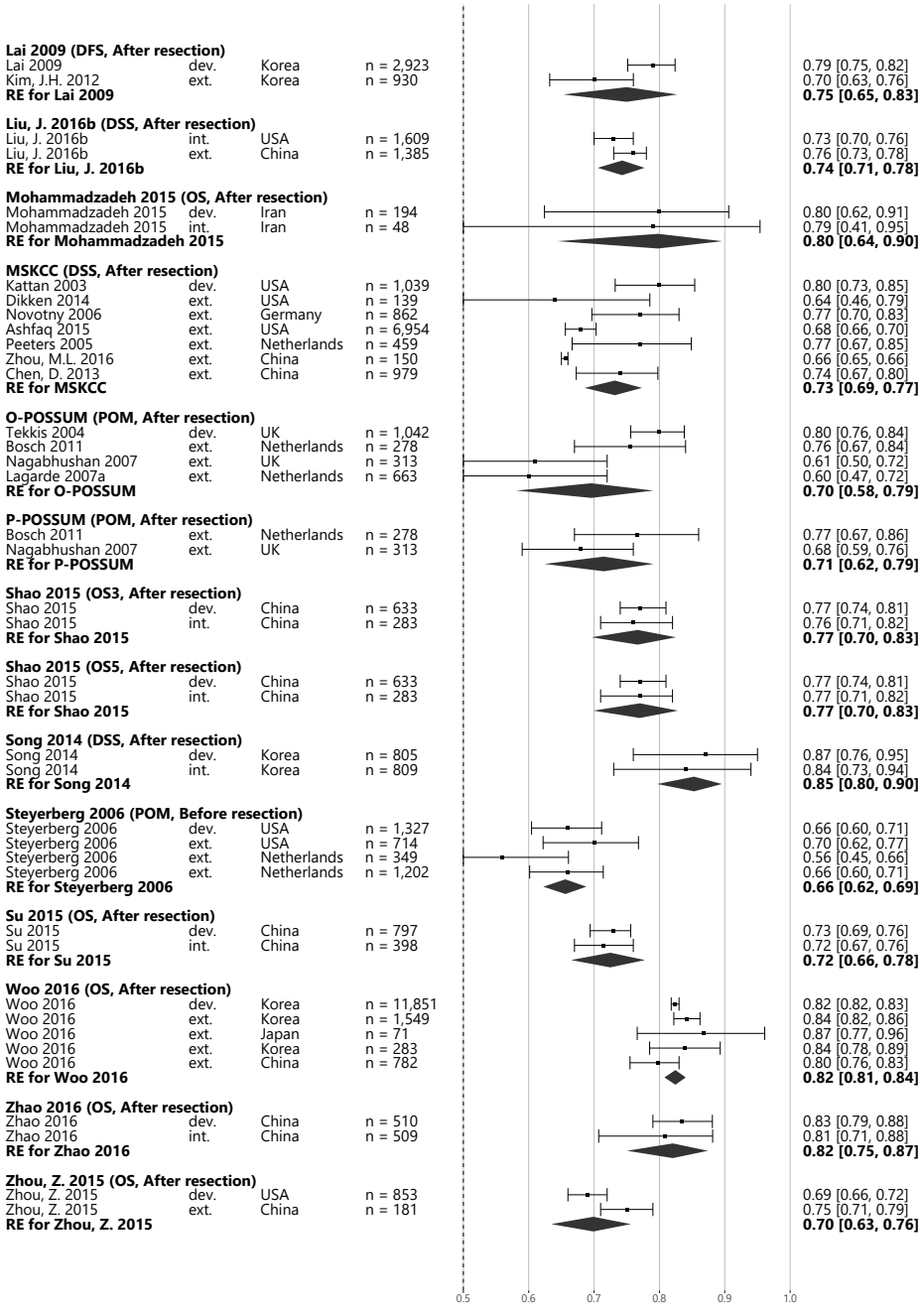


Figure 3B: Random effects meta-analyses of the discriminative abilities (c-indices) of the identified prediction models.

DSS: disease-specific survival; POM: post-operative mortality; OS: overall survival; AE: adverse events; DFS: disease-free survival; REC: cancer recurrence; dev: development c-index, int: internal validation; ext: external validation; RE: random effects meta-analysis estimate.

Discussion

The main aim of this review was to provide an overview of prediction models aimed at predicting survival, adverse events and HRQoL in patients with oesophageal or gastric cancer, and establish their predictive performance and biases.

We identified 45 articles describing the development of novel prediction models and only 16 studies validating these prediction models. We were unable to perform meta-analyses of model calibration, as studies either did not or not adequately report model calibration. The meta-analyses of model discriminative abilities indicate large heterogeneity. The pooled estimates of the discriminative abilities tended to have large confidence intervals, which can be explained by low levels of validation and small cohort sizes. The identified studies generally report a fair discriminative ability for the prediction models. Although nearly every study states that the model is potentially useful in practice, almost all studies do acknowledge the need for further external model validation. However, a mere 10 out of 47 prediction models were subsequently tested in such external validation studies. Indeed, the importance of external validation is shown by the present study as we found that the discriminative ability of models was significantly lower in the validation than in the development phase. Presenting only development results may lead to optimism bias and should be acknowledged when using the prediction models in clinical practice. Large datasets are increasingly being made (freely) available online, which may facilitate more extensive validation of prediction models in the future.

Our findings highlight that prior to using any of these prediction models in clinical practice, clinicians need to carefully consider the number and quality of available validations, the countries/populations in which the models were validated, sample sizes and study biases. In fact, the reported low Reilly and Evans levels of validation indicate that the models we have identified are not ready for widespread implementation in clinical practice. Despite the absence of clinically relevant models, the reported results are essential for future benchmarking and validation studies. Eight models have reached a Reilly and Evans level 3, with the MSKCC model being the most promising with a pooled c-index of 0.73, and extensive validation in a wide-range of populations and settings. We recommend that the MSKCC will be further investigated for its added value in clinical practice in terms of, for example, reduction of decisional conflict and increased patient participation (i.e., shared decision mak-

ing). Only when the quality of care is improved following implementation of the model, its widespread use in clinical practice can be recommended.

Most of the identified models focus on prediction of survival after curative resection of oesophageal or gastric cancer. Although these models provide insight into prognosis of this particular group of patients, they are of limited value for treatment decisions, as treatment has largely been completed at the point of resection. Furthermore, none of the prediction models predict HRQoL, despite the established relevance of HRQoL when making treatment decisions[7], especially in the palliative setting. Finally, in order to make a well-informed treatment choice, patients need to consider both the benefits and harms of treatments to determine which option best fits their preferences and goals. However, none of the prediction models we identified provide estimates of both the benefits and harms associated with a treatment option. Thus, if clinicians opt to use the currently available models, it is imperative that they supplement the information provided by the model with evidence-based predictions concerning not only the possible increase in life-span, but also the possible adverse events and impact on HRQoL.

In order to assess the quality of the studies, we determined sources of possible bias in six different areas. Most studies had a high risk of bias, and all articles showed possible bias in at least one area. The most common bias concerned the handling of missing data. In many studies, it was unclear whether data was missing, how much was missing and how the missing data were handled. Model calibration was not mentioned in some cases and often not accompanied by statistics to provide insight into model quality. Overall, the quality of reporting was poor. Crucial information needed for the interpretation of the results was ill-reported, such as when the model should be used, if the model was to be used with patients for whom treatment has a palliative or curative intent, and what the confidence intervals of the outcomes were. We did not contact authors in cases where the reporting was incomplete, as the focus of this study was to create an overview of reported studies and not to analyse bias in prediction models per se. We strongly advocate that when reporting the development or validation of prediction models the guidance in TRIPOD-statement[15] is followed. This statement provides a checklist of necessary items to include when reporting prediction model development and validation studies, which would facilitate a consistent manner of reporting and safeguard the inclusion of important items needed for interpretation of the data.

In contrast to our expectation, we found no relation between the predictive performance of the models and the impact factor of the journal in which the study is published, nor between the impact factor and study bias. Clinicians should keep in mind that a high impact factor is not a guarantee for quality, and they should always critically assess the quality and generalisability of the prediction model for use in clinical practice. The results of the current study may aid such an evaluation.

In conclusion, we found 47 prediction models intended to predict outcomes in patients with oesophageal and gastric cancer. Most models mainly aimed to predict survival after curative resection. Validation of these models is generally limited and the overall performance was fair. There is a clear need for new prediction models for patients with oesophageal and gastric cancer that focus on both the potential benefits (e.g., improved survival) and harms (e.g., occurrence of adverse events and/or loss of health-related quality of life) of treatment. Such comprehensive prediction models will likely support the decision-making process.

References

1. Ferlay, J., Soerjomataram, I., Ervik, M., *et al.* GLOBOCAN 2012 v1. 0, Cancer incidence and mortality worldwide: IARC CancerBase No. 11. (2013).
2. Nashimoto, A., Akazawa, K., Isobe, Y., *et al.* Gastric cancer treated in 2002 in Japan: 2009 annual report of the JGCA nationwide registry. Gastric Cancer 16, 1–27 (2013).
3. Reim, D., Loos, M., Vogl, F., *et al.* Prognostic implications of the seventh edition of the international union against cancer classification for patients with gastric cancer: The western experience of patients treated in a single-center European institution. Journal of Clinical Oncology 31, 263–271 (2013).
4. SEER. Surveillance, Epidemiology, and End Results: National Cancer Institute. 2013.
5. Jacobs, M., Macefield, R. C., Elbers, R. G., *et al.* Meta-analysis shows clinically relevant and long-lasting deterioration in health-related quality of life after esophageal cancer surgery. Quality of Life Research 23, 1097–1115 (2014).
6. Al-Batran, S. E. & Ajani, J. A. Impact of chemotherapy on quality of life in patients with metastatic esophagogastric cancer. Cancer 116, 2511–2518 (2010).
7. Thrumurthy, S. G., Morris, J. J., Mughal, M. M. & Ward, J. B. Discrete-choice preference comparison between patients and doctors for the surgical management of oesophagogastric cancer. British Journal of Surgery 98, 1124–1131 (2011).
8. Hitz, F., Ribi, K., Li, Q., *et al.* Predictors of satisfaction with treatment decision, decision-making preferences, and main treatment goals in patients with advanced cancer. Supportive Care in Cancer 21, 3085–3093 (2013).
9. Hodgkinson, K., Butow, P., Hunt, G. E., *et al.* The development and evaluation of a measure to assess cancer survivors' unmet supportive care needs: TheCaSUN (Cancer Survivors' Unmet Needs measure). Psycho-Oncology 16, 796–804 (2007).
10. Ravdin, P. M., Siminoff, L. A., Davis, G. J., *et al.* Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. Journal of Clinical Oncology 19, 980–991 (2001).
11. Wishart, G. C., Bajdik, C. D., Azzato, E. M., *et al.* A population-based validation of the prognostic model PREDICT for early breast cancer. European Journal of Surgical Oncology 37, 411–417 (2011).
12. Moons, K. G., de Groot, J. A., Bouwmeester, W., *et al.* Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. PLoS Medicine 11 (2014).
13. Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine 15, 361–387 (1996).
14. Reilly, B. M. & Evans, A. T. Translating clinical research into clinical practice: Impact of using prediction rules to make decisions. Annals of Internal Medicine 144, 201–209 (2006).
15. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. BMC Medicine 13, 148–158 (2015).
16. 2015 Journal Citation Reports®: Clarivate Analytics. 2017. `https://jcr.incites.thomsonreuters.com`.
17. Zhang, H. Z., Jin, G. F. & Shen, H. B. Epidemiologic differences in esophageal cancer between Asian and Western populations. Chinese journal of cancer 31, 281–286 (2012).

18. Kottas, M., Kuss, O. & Zapf, A. A modified Wald interval for the area under the ROC curve (AUC) in diagnostic case-control studies. BMC Medical Research Methodology 14 (2014).
19. Copeland, G. P., Jones, D. & Walters, M. POSSUM: A scoring system for surgical audit. British Journal of Surgery 78, 355–360 (1991).
20. Prytherch, D. R., Whiteley, M. S., Higgins, B., *et al.* POSSUM and Portsmouth POSSUM for predicting mortality. British Journal of Surgery 85, 1217–1220 (1998).
21. Moher, D., Liberati, A., Tetzlaff, J., *et al.* Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Medicine 6 (2009).
22. Biglarian, A., Hajizadeh, E., Kazemnejad, A. & Zali, M. R. Application of artificial neural network in predicting the survival rate of gastric cancer patients. Iranian Journal of Public Health 40, 80–86 (2011).
23. Cao, J., Yuan, P., Wang, L., *et al.* Clinical Nomogram for Predicting Survival of Esophageal Cancer Patients after Esophagectomy. Scientific Reports 6 (2016).
24. Chen, S., Yang, X. & Feng, J. F. A novel inflammation-based prognostic score for patients with esophageal squamous cell carcinoma: The c-reactive protein/ prognostic nutritional index ratio. Oncotarget 7, 62123–62132 (2016).
25. Deans, D. A., Wigmore, S. J., De Beaux, A. C., *et al.* Clinical prognostic scoring system to aid decision-making in gastro-oesophageal cancer. British Journal of Surgery 94, 1501–1508 (2007).
26. Dhir, M., Smith, L. M., Ullrich, F., *et al.* A Preoperative Nomogram to Predict the Risk of Perioperative Mortality Following Gastric Resections for Malignancy. Journal of Gastrointestinal Surgery 16, 2026–2036 (2012).
27. Dikken, J. L., Baser, R. E., Gonen, M., *et al.* Conditional probability of survival nomogram for 1-, 2-, and 3-year survivors after an R0 resection for gastric cancer. Annals of Surgical Oncology 20, 1623–1630 (2013).
28. Duan, J., Deng, T., Ying, G., *et al.* Prognostic nomogram for previously untreated patients with esophageal squamous cell carcinoma after esophagectomy followed by adjuvant chemotherapy. Japanese Journal of Clinical Oncology 46, 336–343 (2016).
29. Eil, R., Diggs, B. S., Wang, S. J., *et al.* Nomogram for predicting the benefit of neoadjuvant chemoradiotherapy for patients with esophageal cancer: A SEER-medicare analysis. Cancer 120, 492–498 (2014).
30. Eom, B. W., Ryu, K. W., Nam, B. H., *et al.* Survival nomogram for curatively resected Korean gastric cancer patients: Multicenter retrospective analysis with external validation. PLoS ONE 10 (2015).
31. Filip, B., Scarpa, M., Cavallin, F., *et al.* Postoperative outcome after oesophagectomy for cancer: Nutritional status is the missing ring in the current prognostic scores. European Journal of Surgical Oncology 41, 787–794 (2015).
32. Fischer, C., Lingsma, H., Hardwick, R., *et al.* Risk adjustment models for short-term outcomes after surgical resection for oesophagogastric cancer. British Journal of Surgery 103, 105–116 (2016).
33. Fuccio, L., Scagliarini, M., Frazzoni, L. & Battaglia, G. Development of a prediction model of adverse events after stent placement for esophageal cancer. Gastrointestinal Endoscopy 83, 746–752 (2016).
34. Gabriel, E., Attwood, K., Shah, R., *et al.* Novel Calculator to Estimate Overall Survival Benefit from Neoadjuvant Chemoradiation in Patients with Esophageal Adenocarcinoma. Journal of the American College of Surgeons 224, 884–894.e1 (2017).
35. Haga, Y., Ikejiri, K., Wada, Y., *et al.* Preliminary study of surgical audit for overall survival following gastric cancer resection. Gastric Cancer 18, 138–146 (2015).

36. Han, D. S., Suh, Y. S., Kong, S. H., *et al.* Nomogram predicting long-term survival after D2 gastrectomy for gastric cancer. Journal of Clinical Oncology 30, 3834–3840 (2012).
37. Hirabayashi, S., Kosugi, S. I., Isobe, Y., *et al.* Development and external validation of a nomogram for overall survival after curative resection in serosa-negative, locally advanced gastric cancer. Annals of Oncology 25, 1179–1184 (2014).
38. Jiang, Y., Zhang, Q., Hu, Y., *et al.* ImmunoScore Signature: A Prognostic and Predictive Tool in Gastric Cancer. Annals of Surgery 267, 504–513 (2016).
39. Jung, H. A., Adenis, A., Lee, J., *et al.* Nomogram to predict treatment outcome of fluoropyrimidine/ platinum-based chemotherapy in metastatic esophageal squamous cell carcinoma. Cancer Research and Treatment 45, 285–294 (2013).
40. Kattan, M. W., Karpeh, M. S., Mazumdar, M. & Brennan, M. F. Postoperative nomogram for disease-specific survival after an R0 resection for gastric carcinoma. Journal of Clinical Oncology 21, 3647–3650 (2003).
41. Kim, Y., Spolverato, G., Ejaz, A., *et al.* A Nomogram to Predict Overall Survival and Disease-Free Survival After Curative Resection of Gastric Adenocarcinoma. Annals of Surgical Oncology 22, 1828–1835 (2015).
42. Kunisaki, C., Miyata, H., Konno, H., *et al.* Modeling preoperative risk factors for potentially lethal morbidities using a nationwide Japanese web-based database of patients undergoing distal gastrectomy for gastric cancer. Gastric Cancer 20, 496–507 (2016).
43. Kurita, N., Miyata, H., Gotoh, M., *et al.* Risk model for distal gastrectomy when treating gastric cancer on the basis of data from 33,917 Japanese patients collected using a nationwide web-based data entry system. Annals of Surgery 262, 295–303 (2015).
44. Lagarde, S. M., Reitsma, J. B., De Castro, S. M., *et al.* Prognostic nomogram for patients undergoing oesophagectomy for adenocarcinoma of the oesophagus or gastro-oesophageal junction. British Journal of Surgery 94, 1361–1368 (2007).
45. Lagarde, S. M., Reitsma, J. B., Maris, A. K. D., *et al.* Preoperative Prediction of the Occurrence and Severity of Complications After Esophagectomy for Cancer With Use of a Nomogram. Annals of Thoracic Surgery 85, 1938–1945 (2008).
46. Lai, J. F., Kim, S., Kim, K., *et al.* Prediction of recurrence of early gastric cancer after curative resection. Annals of Surgical Oncology 16, 1896–1902 (2009).
47. Liu, J., Geng, Q., Chen, S., *et al.* Nomogram based on systemic inflammatory response markers predicting the survival of patients with resectable gastric cancer after D2 gastrectomy. Oncotarget 7, 37556–37565 (2016).
48. Liu, J., Geng, Q., Liu, Z., *et al.* Development and external validation of a prognostic nomogram for gastric cancer using the national cancer registry. Oncotarget 7, 35853–35864 (2016).
49. Liu, J. S., Huang, Y., Yang, X. & Feng, J. F. A nomogram to predict prognostic values of various inflammatory biomarkers in patients with esophageal squamous cell carcinoma. American Journal of Cancer Research 5, 2180–2189 (2015).
50. Marrelli, D., De Stefano, A., De Manzoni, G., *et al.* Prediction of recurrence after radical surgery for gastric cancer: A scoring system obtained from a prospective multicenter study. Annals of Surgery 241, 247–255 (2005).
51. Mohammadzadeh, F., Noorkojuri, H., Pourhoseingholi, M. A., *et al.* Predicting the probability of mortality of gastric cancer patients using decision tree. Irish Journal of Medical Science 184, 277–284 (2015).
52. Muneoka, Y., Akazawa, K., Ishikawa, T., *et al.* Nomogram for 5-year relapse-free survival of a patient with advanced gastric cancer after surgery. International Journal of Surgery 35, 153–159 (2016).

53. Shao, Y., Ning, Z., Chen, J., *et al.* Prognostic nomogram integrated systemic inflammation score for patients with esophageal squamouscell carcinoma undergoing radical esophagectomy. Scientific Reports 5 (2015).
54. Shapiro, J., Van Klaveren, D., Lagarde, S. M., *et al.* Prediction of survival in patients with oesophageal or junctional cancer receiving neoadjuvant chemoradiotherapy and surgery. British Journal of Surgery 103, 1039–1047 (2016).
55. Shiozaki, H., Slack, R. S., Chen, H. C., *et al.* Metastatic Gastroesophageal Adenocarcinoma Patients Treated with Systemic Therapy Followed by Consolidative Local Therapy: A Nomogram Associated with Long-Term Survivors. Oncology (Switzerland) 91, 55–60 (2016).
56. Song, K. Y., Park, Y. G., Jeon, H. M. & Park, C. H. A nomogram for predicting individual survival of patients with gastric cancer who underwent radical surgery with extended lymph node dissection. Gastric Cancer 17, 287–293 (2014).
57. Steyerberg, E. W., Neville, B. A., Koppert, L. B., *et al.* Surgical mortality in patients with esophageal cancer: Development and validation of a simple risk score. Journal of Clinical Oncology 24, 4277–4284 (2006).
58. Su, D., Zhou, X., Chen, Q., *et al.* Prognostic nomogram for thoracic esophageal squamous cell carcinoma after radical esophagectomy. PLoS ONE 10 (2015).
59. Suzuki, A., Xiao, L., Hayashi, Y., *et al.* Nomograms for prognostication of outcome in patients with esophageal and gastroesophageal carcinoma undergoing definitive chemoradiotherapy. Oncology 82, 108–113 (2012).
60. Tekkis, P. P., McCulloch, P., Poloniecki, J. D., *et al.* Risk-adjusted prediction of operative mortality in oesophagogastric surgery with O-POSSUM. British Journal of Surgery 91, 288–295 (2004).
61. Tu, R. H., Lin, J. X., Zheng, C. H., *et al.* Development of a nomogram for predicting the risk of anastomotic leakage after a gastrectomy for gastric cancer. European Journal of Surgical Oncology 43, 485–492 (2017).
62. Woo, Y., Son, T., Song, K., *et al.* A novel prediction model of prognosis after gastrectomy for gastric carcinoma. Annals of Surgery 264, 114–120 (2016).
63. Yang, H. X., Feng, W., Wei, J. C., *et al.* Support vector machine-based nomogram predicts postoperative distant metastasis for patients with oesophageal squamous cell carcinoma. British Journal of Cancer 109, 1109–1116 (2013).
64. Yu, S., Zhang, W., Ni, W., *et al.* Nomogram and recursive partitioning analysis to predict overall survival in patients with stage IIB-III thoracic esophageal squamous cell carcinoma after esophagectomy. Oncotarget 7, 55211–55221 (2016).
65. Zhao, L. Y., Chen, X. L., Wang, Y. G., *et al.* A new predictive model combined of tumor size, lymph nodes count and lymphovascular invasion for survival prognosis in patients with lymph node-negative gastric cancer. Oncotarget 7, 72300–72310 (2016).
66. Zhou, Z., Zhang, H., Xu, Z., *et al.* Nomogram predicted survival of patients with adenocarcinoma of esophagogastric junction. World Journal of Surgical Oncology 13 (2015).
67. Ashfaq, A., Kidwell, J. T., McGhan, L. J., *et al.* Validation of a gastric cancer nomogram using a cancer registry. Journal of Surgical Oncology 112, 377–380 (2015).
68. Bosch, D. J., Pultrum, B. B., De Bock, G. H., *et al.* Comparison of different risk-adjustment models in assessing short-term surgical outcome after transthoracic esophagectomy in patients with esophageal cancer. American Journal of Surgery 202, 303–309 (2011).
69. Chen, D., Jiang, B., Xing, J., *et al.* Validation of the Memorial Sloan-Kettering Cancer Center Nomogram to Predict Disease-Specific Survival after R0 Resection in a Chinese Gastric Cancer Population. PLoS ONE 8 (2013).

70. D'Journo, X. B., Berbis, J., Jougon, J., *et al.* External validation of a risk score in the prediction of the mortality after esophagectomy for cancer. Diseases of the Esophagus 30, 1–8 (2017).
71. Dikken, J. L., Coit, D. G., Baser, R. E., *et al.* Performance of a nomogram predicting disease-specific survival after an R0 resection for gastric cancer in patients receiving postoperative chemoradiation therapy. International Journal of Radiation Oncology Biology Physics 88, 624–629 (2014).
72. Grotenhuis, B. A., Van Hagen, P., Reitsma, J. B., *et al.* Validation of a nomogram predicting complications after esophagectomy for cancer. Annals of Thoracic Surgery 90, 920–925 (2010).
73. Kim, J. H., Kim, H. S., Seo, W. Y., *et al.* External validation of nomogram for the prediction of recurrence after curative resection in early gastric cancer. Annals of Oncology 23, 361–367 (2012).
74. Lagarde, S. M., Maris, A. K., De Castro, S. M., *et al.* Evaluation of O-POSSUM in predicting in-hospital mortality after resection for oesophageal cancer. British Journal of Surgery 94, 1521–1526 (2007).
75. Lagarde, S. M., Reitsma, J. B., Ten Kate, F. J., *et al.* Predicting individual survival after potentially curative esophagectomy for adenocarcinoma of the esophagus or gastroesophageal junction. Annals of Surgery 248, 1006–1012 (2008).
76. Marrelli, D., Morgagni, P., De Manzoni, G., *et al.* External Validation of a Score Predictive of Recurrence after Radical Surgery for Non-Cardia Gastric Cancer: Results of a Follow-Up Study. Journal of the American College of Surgeons 221, 280–290 (2015).
77. Nagabhushan, J. S., Srinath, S., Weir, F., *et al.* Comparison of P-POSSUM and O-POSSUM in predicting mortality after oesophagogastric resections. Postgraduate Medical Journal 83, 355–358 (2007).
78. Novotny, A. R., Schuhmacher, C., Busch, R., *et al.* Predicting individual survival after gastric cancer resection: Validation of a U.S.-derived nomogram at a single high-volume center in Europe. Annals of Surgery 243, 74–81 (2006).
79. Peeters, K. C., Kattan, M. W., Hartgrink, H. H., *et al.* Validation of a nomogram for predicting disease-specific survival after an RO resection for gastric carcinoma. Cancer 103, 702–707 (2005).
80. Reim, D., Novotny, A., Eom, B. W., *et al.* External validation of an eastern asian nomogram for survival prediction after gastric cancer surgery in a European patient cohort. Medicine (United States) 94 (2015).
81. Zafirellis, K. D., Fountoulakis, A., Dolan, K., *et al.* Evaluation of POSSUM in patients with oesophageal cancer undergoing resection. British Journal of Surgery 89, 1150–1155 (2002).
82. Zhou, M. L., Wang, L., Wang, J. Z., *et al.* Validation of the memorial sloan kettering cancer center nomogram to predict disease-specific survival in a chinese gastric cancer population receiving postoperative chemoradiotherapy after an R0 resection. Oncotarget 7, 64757–64765 (2016).

Supplement

Database	Search terms
MEDLINE	(((*Esophageal Neoplasms/ or *Stomach Neoplasms/ or ((esophagus or esophageal or esophagogastric or oesophagus or oesophageal or oesophagogastric or gastroesophag* or gastrooesophag* or gastric or stomach) adj (neoplas* or cancer* or carcino* or adenocarcino* or tumor* or tumour* or malig*)).ti,ab,kw. AND Survival/ or exp Survival Analysis/ or Survival Rate/ or Neoplasm Metastasis/ or Neoplasm Recurrence, Local/ or Comorbidity/ or "Quality of Life"/ or exp Mortality/ or Life Expectancy/ or (surviv* or mortalit* or toxic* or disease free survival or quality of life or QOL or life quality or recurrence or metastas* or comorbidity or life expectanc* or adverse effect* or adverse event*).ti,ab,kw. AND (rule* or scor* or model* or nomogram* or regression* or network* or predict*).ti,ab,kw. AND exp models, statistical/ or exp Regression Analysis/ or Prognosis/ or (validat* or prognos* or predict*).ti,ab,kw.) AND (academic dissertations/ or classical article/ or journal article/ or introductory journal article/ or "review"/)) NOT Animals/ not Humans/) limit to (english language and yr="2000 - 2017")
EMBASE	((exp *esophagus tumor/ or exp *stomach tumor/ or ((esophagus or esophageal or esophagogastric or oesophagus or oesophageal or oesophagogastric or gastroesophag* or gastrooesophag* or gastric or stomach) adj (neoplas* or cancer* or carcino* or adenocarcino* or tumor* or tumour* or malig*)).ti,ab,kw. AND exp survival/ or survival analysis/ or survival rate/ or metastasis/ or tumor recurrence/ or cancer recurrence/ or comorbidity/ or "quality of life"/ or exp mortality/ or life expectancy/ or (surviv* or mortalit* or toxic* or quality of life or QOL or life quality or recurren* or metastas* or comorbidit* or life expectanc* or adverse effect* or adverse event*).ti,ab,kw. AND (rule* or scor* or model* or nomogram* or regression* or network* or predict*).ti,ab,kw. AND statistical model/ or exp regression analysis/ or exp prognosis/ or (validat* or prognos* or predict*).ti,ab,kw.) AND (scientific literature/ or article/ or "review"/)) NOT (animal/ not human/) limit to (english language and yr="2000 - 2017")
PsycINFO	(((esophagus/ or exp Gastrointestinal Disorders/ or exp Gastrointestinal System/ or exp Stomach/) and exp Neoplasms/) or ((esophagus or esophageal or esophagogastric or oesophagus or oesophageal or oesophagogastric or gastroesophag* or gastrooesophag* or gastric or stomach) and (neoplas* or cancer* or carcino* or adenocarcino* or tumor* or tumour* or malig*)).ti,ab,id. AND mortality rate/ or metastasis/ or comorbidity/ or "quality of life"/ or "death and dying"/ or life expectancy/ or (surviv* or mortalit* or toxic* or quality of life or QOL or life quality or recurren* or metastas* or comorbidit* or life expectanc* or adverse effect* or adverse event*).ti,ab,id. AND (rule* or scor* or model* or nomogram* or regression* or network* or predict*).ti,ab,id. AND exp mathematical modeling/ or exp statistical regression/ or prognosis/ or (validat* or prognos* or predict*).ti,ab,id.) limit to human AND (english language and yr="2000 - 2017")

Supplementary Table 1 – *continued from previous page*

Database	Search terms
CINAHL	(MH "Esophageal Neoplasms") OR (MH "Stomach Neoplasms") OR (TI ((esophagus or esophageal or esophagogastric or oesophagus or oesophageal or oesophagogastric or gastroesophag* or gastrooesophag* or gastric or stomach) N1 (neoplas* or cancer* or carcino* or adenocarcino* or tumor* or tumour* or malig*)) OR AB ((esophagus or esophageal or esophagogastric or oesophagus or oesophageal or oesophagogastric or gastroesophag* or gastrooesophag* or gastric or stomach) N1 (neoplas* or cancer* or carcino* or adenocarcino* or tumor* or tumour* or malig*))) AND ((MH "Survival") OR (MH "Survival Analysis+")) OR (MH "Neoplasm Metastasis+") OR (MH "Neoplasm Recurrence, Local") OR (MH "Comorbidity") OR (MH "Quality of Life+") OR (MH "Mortality+") OR (MH "Life Expectancy") OR (TI (surviv* or mortalit* or toxic* or quality of life or QOL or life quality or recurren* or metastas* or comorbidit* or life expectanc* or adverse effect* or adverse event*) OR AB (surviv* or mortalit* or toxic* or quality of life or QOL or life quality or recurren* or metastas* or comorbidit* or life expectanc* or adverse effect* or adverse event*)) AND TI (rule* or scor* or model* or nomogram* or regression* or network* or predict*) OR AB (rule* or scor* or model* or nomogram* or regression* or network* or predict*) AND (MH "Models, Statistical") OR (MH "Regression+") OR (MH "Prognosis+") OR (TI (validat* or prognos* or predict*) OR AB (validat* or prognos* or predict*))
The Cochrane Library	(esophagus or esophageal or esophagogastric or oesophagus or oesophageal or oesophagogastric or gastroesophag* or gastrooesophag* or gastric or stomach) and (neoplas* or cancer* or carcino* or adenocarcino* or tumor* or tumour* or malig*):ti,ab,kw (Word variations have been searched) AND surviv* or mortalit* or toxic* or quality of life or QOL or life quality or recurrence or metastas* or comorbidity or life expectanc* or adverse effect* or adverse event*:ti,ab,kw (Word variations have been searched) AND rule* or scor* or model* or nomogram* or regression* or network* or predict*:ti,ab,kw (Word variations have been searched) AND (MeSH descriptor: [Models, Statistical] explode all trees OR MeSH descriptor: [Regression Analysis] explode all trees OR MeSH descriptor: [Prognosis] explode all trees OR validat* or prognos* or predict*:ti,ab,kw (Word variations have been searched)) Publication Year from 2000 to 2017

Supplementary Table 1: Search strategy per database.

Bias category	Description
Study participant	
++	Prospective study
+	No potential source of bias related to study participants identified High case-mix
?	Not applicable
–	Selection bias; treatment out of date; setting not clearly described; not clearly described whether patients received adjuvant treatment; limited generalisability
– –	Unclear what main treatment modality was; unclear description in- and exclusion criteria; censoring bias; patient characteristics not described
Predictors	
++	Not applicable
+	No potential source of bias related to predictors identified
?	Not applicable
–	Arbitrary cut-off utilised; arbitrary predictor selection; relevant predictors not considered; predictors not defined
– –	Not applicable
Outcome	
++	Not applicable
+	No potential source of bias related to outcome identified
?	Outcome not clearly specified; Number of events not reported
–	Degrees of Freedom/events ratio <10 in development set; timing short-term outcome (e.g., post-operative mortality) unclear
– –	Unclear what main outcome is; timing long-term outcome unclear
Sample size	
++	At least 5,000 cases included in analyses
+	No potential source of bias related to sample size identified; $\geq$ 550 and <5000 cases
?	Not applicable
–	Small sample size ($\geq$ 100 and <550); unsuitable imputation technique
– –	Fewer than 100 cases
Missing data handling	
++	Not applicable
+	No potential source of bias related to missing data handling identified; multiple imputation
?	Unclear which imputation technique was used
–	Unclear how missing data was handled
– –	Complete case analyses; imputation by reference category in nominal variable
Statistical analyses	
++	Not applicable
+	No potential source of bias related to the statistical analyses identified
?	Not applicable
–	Logistic regression for survival (except if outcome was post-operative mortality); confidence intervals were not reported; calibration was not assessed
– –	Discriminatory ability: area under the curve (AUC) nor c-index reported

Supplementary Table 2: Overview and categorisation of potential sources of bias identified in included articles.

Chapter 3

SOURCE: A registry-based prediction model for overall survival in patients with metastatic oesophageal or gastric cancer

Héctor G. van den Boorn, Ameen Abu-Hanna, Emil ter Veer, Jessy Joy van Kleef, Florian Lordick, Michael Stahl, Jaffer A. Ajani, Rosine Guimbaud, Se Hoon Park, Susan J. Dutton, Yung-Jue Bang, Narikazu Boku, Nadia Haj Mohammad, Mirjam A.G. Sprangers, Rob H.A. Verhoeven, Aeilko H. Zwinderman, Martijn G.H. van Oijen and Hanneke W.M. van Laarhoven

Based on:

H.G. van den Boorn, A. Abu-Hanna, E. ter Veer, *et al.* SOURCE: a registry-based prediction model for overall survival in patients with metastatic oesophageal or gastric cancer. Cancers 11 (2019).

Abstract

Prediction models are only sparsely available for metastatic oesophagogastric cancer. Because treatment in this setting is often preference-based, decision-making with the aid of a prediction model is wanted. The aim of this study is to construct a prediction model, called SOURCE, for the overall survival in patients with metastatic oesophagogastric cancer.

Data from patients with metastatic oesophageal (N = 8,010) or gastric (N = 4,763) cancer diagnosed during 2005–2015 were retrieved from the nationwide Netherlands Cancer Registry. A multivariate Cox regression model was created to predict overall survival for various treatments. Predictor selection was performed via the Akaike Information Criterion and a Delphi consensus among experts in palliative oesophagogastric cancer. Validation was performed according to a temporal internal-external scheme. The predictive quality was assessed with the concordance-index (c-index) and calibration.

The model c-indices showed consistent discriminative ability during validation: 0.71 for oesophageal cancer and 0.68 for gastric cancer. The calibration showed an average slope of 1.0 and intercept of 0.0 for both tumour locations, indicating a close agreement between predicted and observed survival. With a fair c-index and good calibration, SOURCE provides a solid foundation for further investigation in clinical practice to determine its added value in shared decision making.

Introduction

Patients with oesophageal or gastric cancer have a relatively poor prognosis. One of the main contributors to the low survival rates is the high prevalence of metastases.[1] Metastatic disease is reported to be present at diagnosis in around 20–30% of oesophageal and in 30–40% of gastric cancer patients.[2,3] Although treatments with curative intent are often not an option when a patient presents with metastatic disease, treatments such as systemic therapy may still prolong life and/or offer symptom relief.[4] Treatment guidelines show, however, that in certain cases best supportive care should be considered in patients with metastatic oesophagogastric cancer.[5,6] As treatment is not always associated with improvement of increased health-related quality of life (HRQoL), the best treatment choice for a particular patient may not be obvious.[7]

Informing patients about their treatment options and the associated risks and benefits can therefore be difficult due to complexity of the patients' disease and heterogeneity of outcomes.[8] Prediction models, however, can aid in this process and allow individualised decision making.[9] Over the years various prediction models have been developed to support this process, by predicting outcomes such as survival and recurrence in cancer patients. The *Adjuvant! Online* prediction model, for example, predicts survival in breast cancer patients on the basis of various demographic and clinical variables.[10] An important feature is the comparison of various treatments by displaying the added survival benefit. Recently a review of the prediction models for oesophageal and gastric cancer showed that nearly all prediction models available for oesophagogastric cancer are aimed at predicting survival after curative treatment.[11] Only two prediction models are available that are intended for patients with metastatic disease. The model by Jung et al.[12] predicts overall survival based on a dataset of 239 South Korean patients with oesophageal squamous cell carcinoma. All patients were treated with either fluorouracil/cisplatin or capecitabine/cisplatin in a first-line setting. The model was presented as a nomogram and predicts the one-year survival probability. For the model by Shiozaki et al.[13], 64 patients with metastatic adenocarcinomas were included and all received chemotherapy followed by chemoradiation. The model, intended for patients with favourable outcomes, was also presented as a nomogram and predicts the median overall survival time.

Given the restrictive inclusion criteria and small sample sizes, the generalisability of these models is likely to be limited which possibly hampers implementation in clinical practice. A model is needed that focuses on patients with metastatic disease and informs on the various treatment options which the patient is facing.

It is therefore the aim of this study to create and evaluate a prediction model based on a large nationwide dataset for use in clinical practice, called SOURCE (Stimulating evidence-based, personalised and tailored information provision to improve decision making after Oesophagogastric Cancer diagnosis). SOURCE is intended to predict overall survival for a variety of treatment options in a heterogeneous group of patients with metastatic oesophageal or gastric cancer.

Methods

This report was written in accordance to the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) guidelines.[14] Data of the prospectively maintained population-based Netherlands Cancer Registry (NCR) was used in the development and validation of the prediction model. The records of all 14,422 metastatic oesophageal and gastric cancer patients diagnosed between January 2005 and December 2015 were retrieved from the NCR. Patients with unknown follow-up (N = 4), patients who had cT_0 tumours (N = 5) and patients with cancer types other than carcinomas (N = 227) were excluded from further analysis. Additionally, patients who died within 14 days after diagnosis (N = 697) were also excluded, because patients in such poor health would not likely use a prediction model. Patients with multiple primary tumours (N = 9) retained their initial tumour in the dataset, and subsequent tumours were excluded. Finally, patients whose only distant metastases were located in lymph nodes in the head or neck region, were excluded from further analyses (N = 707). These patients could be treated with a curative intent and therefore fall outside the scope of the prediction model. This left a total of 8,010 oesophageal cancer patients and 4,763 gastric cancer patients for inclusion in the dataset.

The outcome of the SOURCE prediction model is overall survival as it gives the most complete survival information for patients. It was measured from the date of diagnosis to the date of death, or the date of last follow up when the patient was censored.
The development of the SOURCE model consisted of three high-level steps which are explained below. First, multiple prediction models were built using Cox regression models. The models were validated in patients diagnosed in a single year and were constructed based on records from previous years. For example, records from patients diagnosed in 2012 were used to validate a prediction model based on records from patients diagnosed up to 2012 (i.e., 2005 through 2011). This was repeated for each validation cohort and therefore a total of 10 prediction models were constructed. Second, the validation results for these 10 models were meta-analysed to investigate the model overfit. Third, the final SOURCE prediction model was created based on the complete dataset.

Predictor selection and Delphi consensus

A set of possible predictors in the NCR dataset was established. Variables with more than 50% missing values, variables with the same value for all patients (which are therefore non-informative) and nominal variables with less than 50 cases for each category were discarded from the NCR dataset. All other variables remained as possible model predictors.

A modified two-round Delphi consensus, similar to the COMM-PACT study[15] in metastatic pancreatic cancer, was performed to extend this set with possibly important predictors that were missed. A systematic review on prognostic factors in advanced oesophagogastric cancer served as a basis for the Delphi consensus procedure.[16] All corresponding authors of the 41 phase III trials included in the systematic review were invited to participate in this study. During the first round, the experts received a list of 56 possible predictors of overall survival in metastatic oesophagogastric cancer, obtained from the systematic review.[16] For each predictor the number of studies investigating its effect and the estimated effect sizes were given. The experts were free to select as many predictors of overall survival as they deemed necessary, stratified by tumour location and treatment if needed, and were given the opportunity to include additional predictors.

After the first Delphi round, all predictors that were selected by at least 50% of the experts were included in the consensus list. Predictors selected by 20% to 50% of the experts and additional predictors that were suggested by the experts, were presented during the second consensus round alongside the results of the first round. Again, predictors selected by at least 50% of the experts in the second round were included in the final consensus list. Subsequently, all selected predictors on the consensus list were added to the set of possible predictors if available in the NCR dataset or if the predictors could be derived from other variables.

The set of possible NCR variables and predictors selected by the experts in the Delphi consensus formed the initial set of predictors. During the model specifications, predictors were selected from this joint set.

Development and validation of the prediction model

For the development of the prediction model, a Cox proportional hazard model with overall survival as the main outcome was developed using the *rms* package in the R-studio environment with R version 3.3.4.[17–19] An overview of the model development process is shown in Figure 1.

To increase model generalisability and robustness, an internal-external temporal cross-validation was employed.[20] With this scheme, the data were split into so-called folds according to the patient diagnosis year. For each fold, the model was evaluated on a patient cohort diagnosed in a single year and the model was constructed on the data of all patients from earlier diagnosis years, thus mimicking a true external temporal validation. Within each fold, multiple imputation (m = 5) by chained equations was used to handle missing data.[21] Conditional multiple imputation was employed to transform TNM-variables from the sixth edition used for patients diagnosed prior to 2010, to the seventh edition used for patients diagnosed as of 2010.[22,23] Specifically, these transformations were as follows: for oesophageal cancer, cN_1 was transformed into $cN_1/cN_2/cN_3$ and cM_{1A} was transformed to $cN_1/cN_2/cN_3$ and cM_0. For gastric cancer, cN_1 was transformed to cN_2/cN_3 and cT_2 into cT_2/cT_3. With these transformations, the meaning of the cTNM variables remained consistent across the entire dataset, while the uncertainty of the transformations was taken into account by multiple imputation. For each fold, bidirectional selection was performed using the AIC procedure to select from the initial predictor set including the predictors suggested during Delphi procedure.[24] Interactions between the predictor set and 'initial treatment' were subsequently added if the AIC statistic improved. Due to the stochastic nature of multiple imputations, the predictor selections could differ in each of the five multiple imputation rounds. Predictor pooling therefore took place by including predictors only if they were selected in at least three out of five multiple imputation rounds.

The Cox regression models were subsequently constructed for each imputation using the selected predictors. The concordance-index (c-index), calibration slope, intercept and deviance measured the model's performance and were obtained for both the development and validation cohorts. The c-index is a measure of discrimination and ranges from 0.5 (no discrimination at all) to 1 (perfect discrimination).[25] Calibration measures the goodness-of-fit and is described by the agreement between predicted and observed outcomes at the median overall survival time (5.1 months for oesophageal cancer and 3.9 months for gastric cancer).[26] A linear model is used to describe this congruence and has an intercept of 0 and slope of 1 when the predictions are perfect.[26] The calibration deviance is determined by the average absolute deviance between the predicted and observed survival.[27]

Finally, the performance results were pooled across all five imputations for each fold. The Cox regression models were combined into a single prediction model with pooled parameter values. The performance measures were subsequently meta-analysed with a random-effects model across all folds to obtain the internal-external validation scores. The performance measures were calculated on data in the validation cohort as well as the full model, thus an estimation of the model overfit can be made.

The construction of the full SOURCE prediction model followed identical steps. However, the complete dataset was used to construct and validate the model and the data were therefore not split into folds.

Research ethics

According to the Central Committee on Research involving Human Subjects, this type of study does not require approval from an ethics committee in the Netherlands. However, the study was approved by the Privacy Review Board of the Netherlands Cancer Registry (project code K17134).

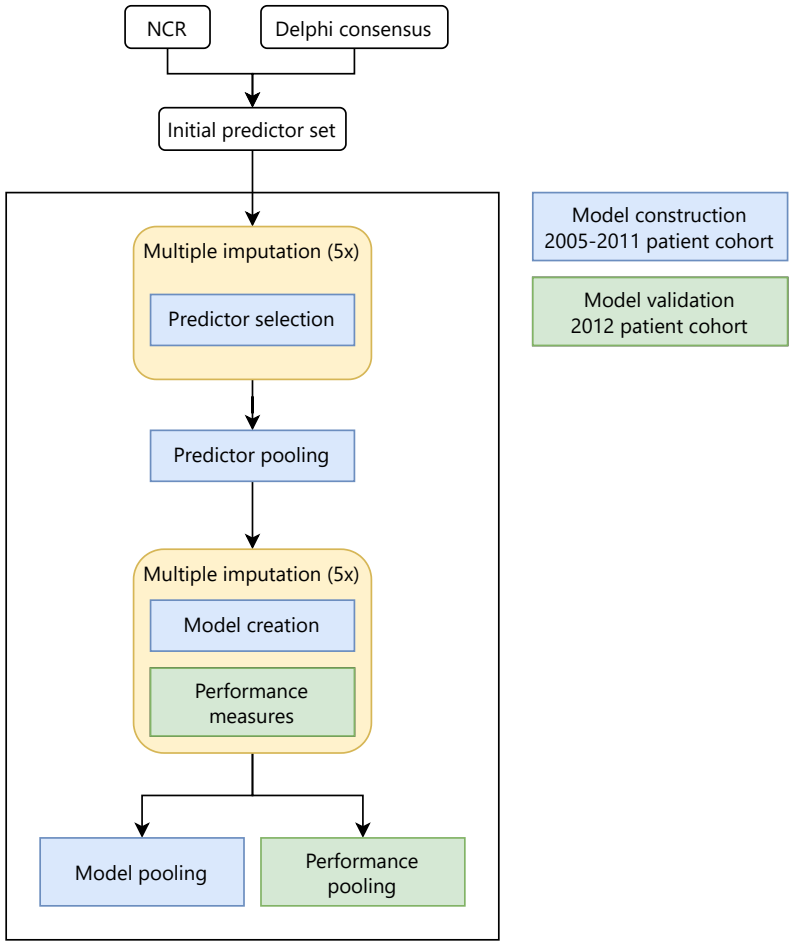


Figure 1: Example model creation and validation.
The figure shows the construction and validation of a prediction model. This method was used during temporal cross-validation and construction of the final model. The image illustrates in this particular case the model construction based on the 2005–2011 patient cohort (shown in blue) and validation in the 2012 patient cohort (shown in green). An initial predictor set is created with variables from the NCR and extended with predictors from the Delphi consensus. We used multiple imputation for the handling of missing data after which predictors were selected by the bidirectional Akaike Information Criterion (AIC) procedure. Since the predictors selected by the AIC procedure may differ in each imputation, the model predictors were pooled by selecting the predictors occurring in the majority of imputations (in at least three out of five imputations). For each imputation, a model was created and validated on the 2012 patient cohort. The model parameters were pooled to establish the model for this cohort, and likewise the performance measures were pooled. This procedure was employed for all internal-external temporal validations; the model was validated on a patient cohort diagnosed in a single year and constructed on a patient cohort of all patients diagnosed in earlier years. For the final SOURCE model, the complete cohort is used for construction and validation of the model.

Results

An overview of the metastatic oesophageal (N = 8,010) and gastric cancer (N = 4,763) patients whose data were used to create the prediction model, is given in Table 1.

Selected predictors

Of the corresponding authors of 41 phase III trials who were invited, eight agreed to participate in the Delphi consensus and completed both rounds.[16] In round one, 14 of the 56 predictors were retained and 25 were excluded. Additionally, 17 predictors were selected by 20–50% of the experts, and eight new predictors were proposed by the experts. These 25 predictors were considered during the second consensus round. Finally, three predictors were included during the second consensus round. The total number of included predictors of the first and second round therefore is 17. The outcomes of the Delphi procedure are displayed in more detail in Supplementary Table 1 and the final selection of the predictors determined by the consensus are displayed in Table 2 alongside the selected SOURCE predictors. The Delphi consensus procedure selected nine predictors which were unavailable in the NCR dataset and could therefore not be included in the list of preselected variables. Seven predictors selected in the Delphi consensus were available in the NCR, all of which were selected as predictors in the final SOURCE models.

Variable	Oesophagus	Gastric
N (deaths)	8,010 (7,825)	4,763 (4,673)
Median overall survival in months (IQR)	5.1 (2.2–10.1)	3.9 (1.7–8.4)
Age (mean (sd))	66.80 (10.91)	68.58 (12.34)
Sex (%)		
Male	6,284 (78.5)	2,858 (60.0)
Female	1,726 (21.5)	1,905 (40.0)
cT stage (%)		
Missing	1 (0.0)	1 (0.0)
cT_1	108 (1.3)	58 (1.2)
cT_2	1,388 (17.3)	659 (13.8)*
cT_3	1,822 (22.7)	672 (14.1)*
cT_4	694 (8.7)	802 (16.8)
cT_X	3,997 (49.9)	2,571 (54.0)
cN stage (%)		
Missing	1 (0.0)	0 (0.0)
cN_0	2,127 (26.6)	2,366 (49.7)
cN_1	2,502 (31.2)*	1,012 (21.7)*
cN_2	2,391 (29.9)*	1,264 (27.0)*
cN_3	989 (12.3)*	121 (2.5)
Primary oespohageal tumour topography (%)		
Cervical	44 (0.5)	
Upper thoracic	205 (2.6)	
Mid-thoracic	713 (8.9)	
Lower thoracic	4,461 (55.7)	
Overlapping lesion	315 (3.9)	
Junction	2,112 (26.4)	
NOS	160 (2.0)	
Primary gastric tumour topography (%)		
Fundus		162 (3.4)
Corpus		954 (20.0)
Antrum Pylori		1,075 (22.6)
Pylorus		239 (5.0)
Lesser curvature NOS		181 (3.8)
Greater curvature NOS		106 (2.2)
Overlapping lesion		1,645 (34.5)
NOS		401 (8.4)
Histological type (%)		
Adenocarcinoma	6,321 (78.9)	4,691 (98.5)
Squamous cell	1,423 (17.8)	0 (0.0)
Other	266 (3.3)	72 (1.5)
Differentiation grade (%)		
Missing	3,472 (43.3)	2,180 (45.8)
G1	112 (1.4)	42 (0.9)
G2	1,464 (18.3)	488 (10.2)
G3	2,896 (36.2)	2,028 (42.6)
G4	66 (0.8)	25 (0.5)
Only distance lymphnode metastasis (%)		
Missing	267 (3.3)	168 (3.5)
No	6,532 (81.5)	4,141 (86.9)
Yes	1,211 (15.1)	454 (9.5)
Liver metastasis (%)		
Missing	267 (3.3)	168 (3.5)
No	3,699 (46.2)	2,873 (60.3)
Yes	4,044 (50.5)	1,722 (36.2)

Table 1 – *continued from previous page*

Variable	Oesophagus	Gastric
Peritoneal metastasis (%)		
Missing	267 (3.3)	168 (3.5)
No	7,190 (89.8)	2,735 (57.4)
Yes	553 (6.9)	1,860 (39.1)
Number of metastatic sites (%)		
Missing	267 (3.3)	168 (3.5)
1	4,457 (55.6)	3,099 (65.1)
2	2,208 (27.6)	1,067 (22.4)
≥ 3	1,078 (13.5)	429 (9.0)
Initial treatment (%)		
None	2,131 (26.6)	2,266 (47.6)
Chemotherapy	2,216 (27.7)	1,648 (34.6)
Radiotherapy (primary tumour)	2,081 (26.0)	154 (3.2)
Radiotherapy (metastasis)	367 (4.6)	63 (1.3)
Chemoradiation	80 (1.0)	0 (0.0)
Chemotherapy + short-term radiation	317 (4.0)	52 (1.1)
Resection (primary tumour)	0 (0.0)	247 (5.2)
Resection (metastasis)	56 (0.7)	97 (2.0)
Stent	298 (3.7)	56 (1.2)
Other	464 (5.8)	180 (3.8)

Table 1: Overview of patient characteristics stratified per tumour location.
cT stage, cN stage and differentiation grade defined are according to the TNM staging system, 7th edition. NOS: Not otherwise specified; CI: 95% confidence interval; IQR: Inter-quarter range; SD: Standard deviation; *: Conditional variable imputation (see Methods), these patients had non-missing TNM 6th variables which were transformed to the indicated TNM 7th edition stages.

Predictor	Delphi consensus	Oesophagus model	Gastric model
Age	x	x	x
Sex			x
cT stage		x	x
cN stage		x	x
Topography of primary tumour	x	x	
Histological type	x	x	
Tumour differentiation grade		x	x
Lymph node metastasis in head/neck area		x	
Intra-thoracic lymph node metastasis			x
Intra-abdominal lymph node metastasis		x	x
Only distant lymph node metastasis		x	x
Liver metastases	x	x	
Peritoneal metastases	x	x	
Number of metastatic sites	x	x	
Initial treatment	x	x	x
Peritoneal metastases with ascites	x		
Performance status	x		
Histology (Lauren)	x		
Weight loss	x		
Tumour Microsatellite Instability (MSI) status	x		
Region/country	x		
HER status	x		
Disease status (unresectable vs. recurrent)	x		
Bilirubin	x		

Table 2: List of the prediction model predictors.
The variables selected by the experts are shown in the left column and variables selected for the final prediction models in the middle and right columns. Predictors indicated in bold were available in the Netherlands Cancer Registry (NCR) dataset and could be used for the creation of the SOURCE prediction model.

Final model parameters

The model parameters of the resulting SOURCE model for overall survival in metastatic oesophageal cancer and metastatic gastric cancer are presented in Table 3.

The performance measures for both the complete SOURCE model and the internal-external validation are shown in Table 4. The results show that the prediction model has a slightly better performance in oesophageal cancer than in gastric cancer. The calibration slopes and intercepts lie close to the optimal values of 1 and 0, respectively. While the performance measures are marginally lower during validation than in the complete model, the correspondence between both settings remains high. The meta-analyses of the model performance statistics are shown in Figure 2 for oesophageal cancer and in Figure 3 for gastric cancer. These figures show the performance statistics for each validation cohort. The calibration plots of the temporal validation cross-validations are shown in Figure 4.

Covariate	Oesophagus HR (CI)	Stomach HR (CI)
Age	1.001 (0.996–1.005)	1.003 (0.999–1.007)
Sex		
Male		1
Female		0.953 (0.898–1.012)
cT stage		
cT_1	1	1
cT_2	1.204 (0.983–1.474)	0.928 (0.704–1.223)
cT_3	1.103 (0.901–1.349)	0.856 (0.650–1.128)
cT_4	1.459 (1.182–1.800)	0.995 (0.756–1.309)
cT_X	1.459 (1.197–1.777)	1.013 (0.775–1.324)
cN stage		
cN_0	1	1
cN_1	0.974 (0.918–1.034)	0.900 (0.834–0.971)
cN_2	1.030 (0.969–1.096)	0.996 (0.927–1.071)
cN_3	1.154 (1.061–1.255)	0.957 (0.793–1.156)
Tumour topography		
Cervical	1	
Upper thoracic	1.039 (0.744–1.450)	
Mid-thoracic	0.989 (0.723–1.351)	
Lower thoracic	1.062 (0.779–1.447)	
Overlapping lesion	1.226 (0.886–1.697)	
Junction	0.999 (0.730–1.367)	
NOS	1.181 (0.837–1.665)	
Histological type		
Adenocarcinoma	1	
Squamous cell	1.011 (0.942–1.085)	
Other	1.168 (1.005–1.358)	
Differentiation grade		
G1	1	1
G2	0.949 (0.825–1.090)	1.294 (1.049–1.596)
G3	1.124 (0.981–1.288)	1.524 (1.245–1.865)
G4	1.396 (1.051–1.854)	1.734 (1.223–2.459)
Lymph node metastasis in head/neck area		
No	1	
Yes	0.868 (0.790–0.954)	
Intra-thoracic lymph node metastasis		
No	1	1
Yes	0.548 (0.430–0.698)	0.739 (0.628–0.870)
Intra-abdominal lymph node metastasis		
No	1	1
Yes	0.834 (0.742–0.938)	0.902 (0.811–1.003)
Only distant lymph node metastasis		
No	1	1
Yes	0.788 (0.732–0.849)	0.771 (0.694–0.856)
Liver metastasis		
No	1	
Yes	1.222 (1.156–1.292)	
Peritoneal metastasis		
No	1	
Yes	1.274 (1.158–1.401)	
Number of metastatic sites	1.347 (1.270–1.429)	1.335 (1.247–1.430)
Initial treatment (IT)		
None	1	1
Chemotherapy	0.237 (0.151–0.372)	0.436 (0.287–0.664)
Radiotherapy (primary tumour)	0.238 (0.151–0.375)	1.428 (0.363–5.619)
Radiotherapy (metastasis)	0.386 (0.169–0.884)	8.419 (1.754–40.411)

Table 3 – *continued from previous page*

Covariate	Oesophagus HR (CI)	Stomach HR (CI)
Chemoradiation	0.246 (0.042–1.455)	
Chemotherapy + short-term radiation	0.280 (0.110–0.715)	1.268 (0.138–11.611)
Resection (primary tumour)		0.427 (0.169–1.080)
Resection (metastasis)	0.029 (0.004–0.227)	0.092 (0.027–0.313)
Stent	0.881 (0.313–2.478)	1.441 (0.132–15.795)
Other	0.121 (0.058–0.250)	0.422 (0.143–1.250)
IT = Chemotherapy		
**Intra-thoracic lymph node metastasis*	1.798 (1.255–2.577)	
**Intra-abdominal lymph node metastasis*	1.091 (0.935–1.275)	
**Age*	1.005 (0.999–1.011)	1.000 (0.994–1.006)
**Number of metastatic sites*	0.825 (0.760–0.895)	0.864 (0.786–0.949)
IT = Radiotherapy (primary tumour)		
**Intra-thoracic lymph node metastasis*	1.481 (1.080–2.031)	
**Intra-abdominal lymph node metastasis*	1.266 (1.086–1.476)	
**Age*	1.009 (1.003–1.015)	0.990 (0.974–1.007)
**Number of metastatic sites*	0.910 (0.836–0.990)	0.918 (0.681–1.239)
IT = Radiotherapy (metastasis)		
**Intra-thoracic lymph node metastasis*	0.972 (0.354–2.668)	
**Intra-abdominal lymph node metastasis*	1.432 (0.963–2.130)	
**Age*	1.009 (0.997–1.020)	0.976 (0.958–0.995)
**Number of metastatic sites*	0.901 (0.790–1.028)	0.706 (0.516–0.965)
IT = Chemoradiation		
**Intra-thoracic lymph node metastasis*	4.522 (0.594–34.393)	
**Intra-abdominal lymph node metastasis*	4.407 (0.588–33.038)	
**Age*	1.005 (0.981–1.031)	
**Number of metastatic sites*	0.746 (0.459–1.212)	
IT = Chemotherapy + short-term radiation		
**Intra-thoracic lymph node metastasis*	0.940 (0.495–1.784)	
**Intra-abdominal lymph node metastasis*	0.921 (0.689–1.231)	
**Age*	1.004 (0.991–1.018)	0.990 (0.962–1.020)
**Number of metastatic sites*	0.819 (0.706–0.949)	0.717 (0.507–1.015)
IT = Resection (primary)		
**Age*		0.999 (0.987–1.011)
**Number of metastatic sites*		0.955 (0.717–1.271)
IT = Resection (metastasis)		
**Intra-thoracic lymph node metastasis*	7.155 (0.947–53.490)	
**Intra-abdominal lymph node metastasis*	1.089 (0.385–3.084)	
**Age*	1.038 (1.005–1.071)	1.025 (1.009–1.042)
**Number of metastatic sites*	0.810 (0.541–1.213)	0.879 (0.668–1.156)
IT = Stent		
**Intra-thoracic lymph node metastasis*	2.640 (1.175–5.931)	
**Intra-abdominal lymph node metastasis*	1.027 (0.737–1.430)	
**Age*	1.001 (0.988–1.014)	0.997 (0.968–1.027)
**Number of metastatic sites*	1.025 (0.871–1.206)	0.957 (0.656–1.396)
IT = Other		
**Intra-thoracic lymph node metastasis*	1.195 (0.623–2.291)	
**Intra-abdominal lymph node metastasis*	0.889 (0.685–1.153)	
**Age*	1.019 (1.009–1.029)	1.012 (0.998–1.027)
**Number of metastatic site*	1.229 (1.056–1.431)	0.803 (0.625–1.032)

Table 3: Prediction model for overall survival in patients with metastatic oesophageal and gastric cancer.
Initial treatment interactions terms are given in italics. HR: Hazard Ratio. NOS: Not otherwise specified. CI: 95% confidence interval. IT: Initial treatment.

	Oesophageal cancer		Gastric cancer	
	Complete model	**Internal-external validation**	**Complete model**	**Internal-external validation**
C-index	0.71 (0.71–0.72)	0.71 (0.70–0.71)	0.69 (0.68–0.70)	0.68 (0.67–0.69)
Calibration slope	1.01 (1.01–1.01)	1.02 (0.96–1.07)	0.99 (0.99–0.99)	1.01 (0.89–1.13)
Calibration intercept	-0.00 (-0.00– -0.00)	-0.02 (-0.05–0.01)	-0.01 (-0.01– -0.01)	-0.01 (-0.06–0.04)
Calibration deviance	0.00 (0.00–0.00)	0.02 (0.01–0.04)	0.01 (0.01–0.01)	0.03 (0.02–0.04)

Table 4: Performance measures for the SOURCE in oesophagus and gastric cancer.

The discrimination index and calibration statistics are shown side-by-side for both the complete SOURCE model as well as for the internal-external temporal validation. The 95% confidence interval is stated in parentheses for each outcome.

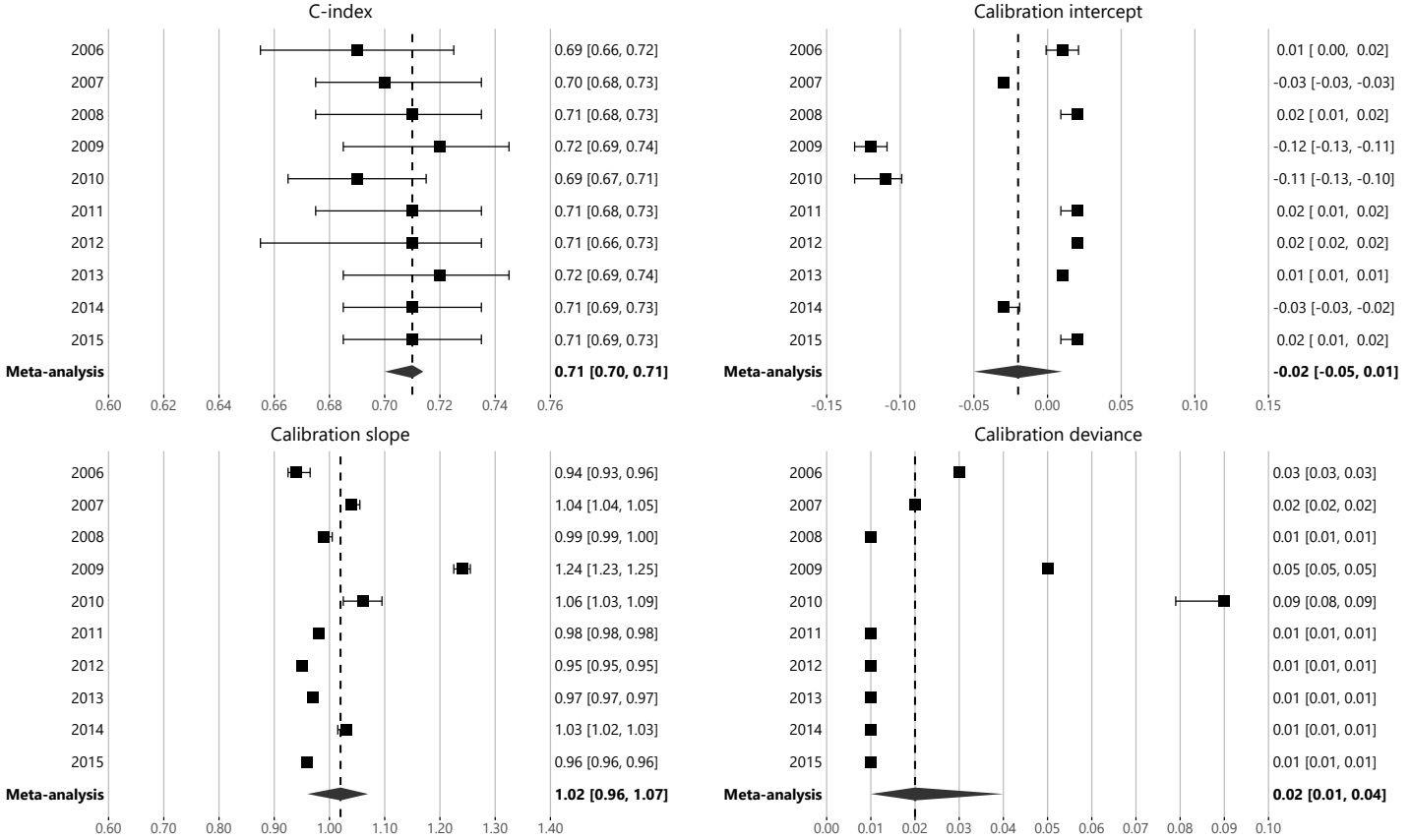


Figure 2: Meta-analysis for internal-external cross-validation (oesophagus).
Each of the four panels shows the meta-analysis of the model outcomes for oesophageal cancer patients. The year indicates on which diagnosis year cohort the model is validated.

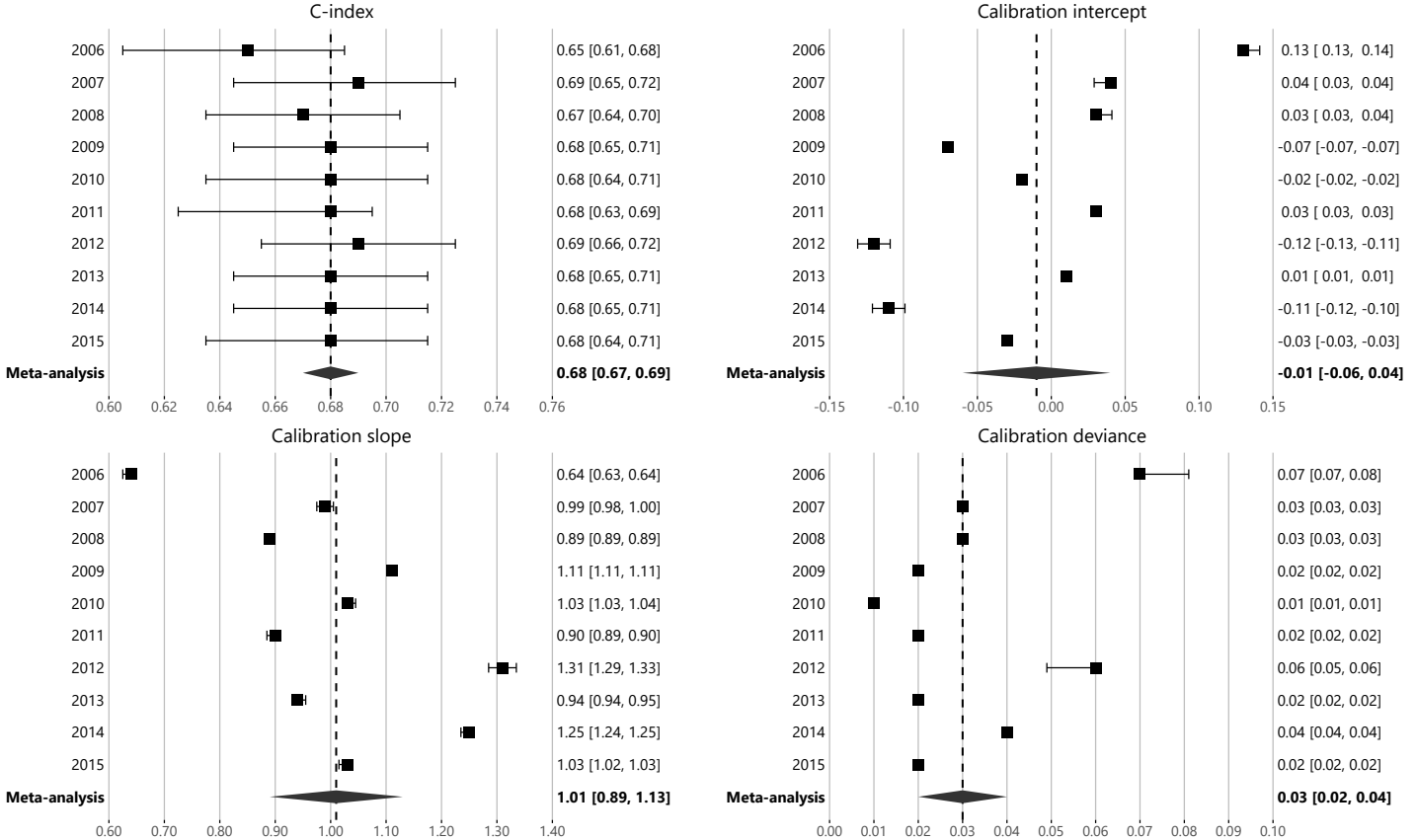


Figure 3: Meta-analysis for internal-external cross-validation (stomach).
Each of the four panels shows the meta-analysis of the model outcomes for gastric cancer patients. The year indicates on which diagnosis year cohort the model is validated.

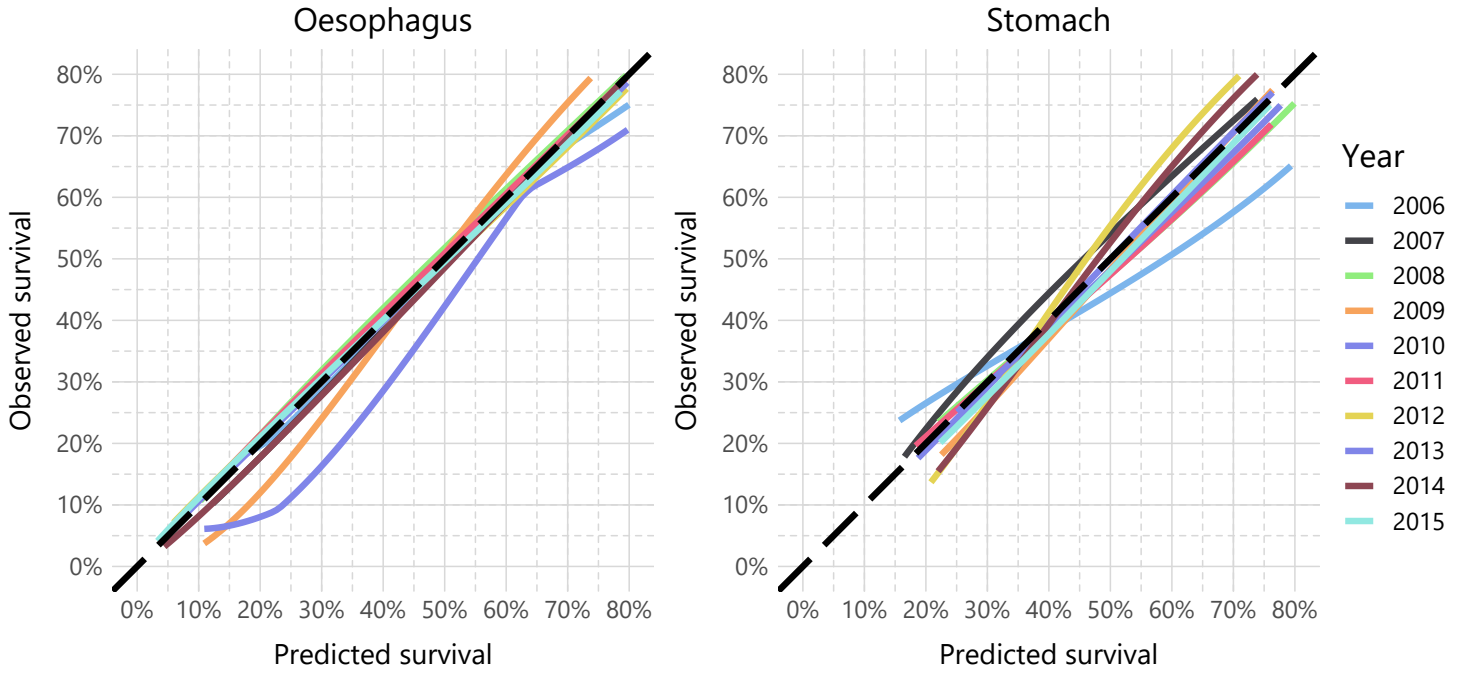


Figure 4: Calibration plots during temporal cross-validation for the oesophageal model (left) and the gastric model (right).

The different lines indicate the correspondence between predicted and observed survival for various diagnosis years. The calibration plot was established at the median overall survival (5.1 months for oesophageal cancer and 3.9 months for gastric cancer). The dashed line indicates an ideal calibration line with an intercept of 0 and slope of 1.

Discussion

The SOURCE model presented in this paper is the first prediction model for survival outcome of metastatic oesophageal and gastric cancer patients that was created with a large (N = 12,773) nation-wide cohort, and includes treatment as a separate predictor. This allows for a flexible model enabling the provision of prognoses for various treatments and tumour locations within the upper gastrointestinal tract. Importantly, the predictors included in the SOURCE model are available in standard clinical care and do not require additional tests that may be cost prohibitive. The strengths of SOURCE lie in its clinical applicability, providing a model for all metastatic oesophagogastric cancer patients and including various treatment options.

In creating SOURCE, various steps were undertaken to increase the quality of the model, its reproducibility and its robustness. First, the predictors were selected both by the bidirectional Akaike Information Criterion (AIC) procedure and a Delphi consensus procedure, thereby combining "the best of both worlds" including data-driven analysis and expert clinician-guided selection. Although not all proposed predictors were available in the current dataset, future models can be built using this selection. Secondly, a temporal internal-external cross-validation method was employed based on the year of diagnosis.[20] With this approach, advances in patient care and treatment are taken into account. In addition, this approach is comparable to a true external temporal validation where an existing prediction model is validated on new patients. Lastly, instead of a complete case analysis, which excludes patients with missing data and thereby could increase bias, we employed the robust multiple imputation method for handling missing data.[28] This not only has the advantage of dealing with uncertainty of the imputations, but can also be used for the transformation of specific variables, such as TNM staging, thus enabling a richer dataset on which the prediction model was based. With these methods, it is possible to obtain a more precise estimate of the model parameters while keeping the amount of overfit small.

Indeed, the SOURCE model showed a fair discriminative ability, with a c-index of approximately 0.71 for oesophageal cancer and 0.68 for gastric cancer. Although certain other models were able to discriminate better between patients, it must be noted that our dataset was relatively homogeneous, including patients with metastatic oesophagogastric cancer only.[29,30] Differentiating between survival outcomes of a rather homogeneous group of patients is more complex than differentiating between survival outcomes of patients

with cancers from various primary origins and known large differences in survival. The model further shows an overall good average accordance between predicted and observed survival. These results remain consistent between the full model and the internal-external temporal cross-validation, thus indicating a lack of overfit. Additional external validation with cohorts from other countries and more recent years is needed to further examine the robustness of the model.

Some limitations of this study have to be acknowledged. First of all, nine predictors selected in the Delphi consensus were not represented in the dataset (see Table 2) and could therefore not be included in the final prediction model. Inclusion of at least some of these predictors, would likely have improved the model's performance. One of the most important predictors of overall survival, performance status, was reported limitedly in the NCR only for the year 2015 and could therefore not be included in the analysis.[31] Further, the initial treatment variable lacks detail. Ideally, several therapies would be subdivided (such as various chemotherapy regimens) to enable a better fit of the model parameters. However, performance status and more detailed treatment information will be available more abundantly in future years and could become predictors in the prediction model. This stresses the need for intermittent updating of predictions models when new data becomes available to increase the model's performance and keep up with the development of new treatment options over time. It can also be noted that the models display in some cases hazard ratio's below 1 for cT and cN stages, implying an unexpected slight decrease of hazard compared to the cT_1 and cN_1 stages. We hypothesise that this may be caused by aggressive tumour behaviour resulting in a shorter overall survival in patients who developed metastases despite a low cT or cN stage. Lastly, SOURCE predicts overall survival at diagnosis. However, due to the nature of the registration process the dataset also erroneously included patients initially diagnosed as cM_0 but whose staging was updated within six weeks to cM_1 due to disease progression or the discovery of metastases during additional diagnostic testing. Consequently, patients who started treatment with curative intent, such as resections, may be overrepresented. Unfortunately, these patients could not be identified and excluded. Based on a detailed analysis of a subset of patients, we estimate this percentage to be small ($\sim 6\%$).

Use of the SOURCE model could be valuable and helpful in clinical practice and stimulate shared decision making. In shared decision making, well balanced provision of information is key.[32] The possibility to compare different treatment options, e.g., chemotherapy and best supportive care could stimulate shared decision making. Figure 5 shows how the SOURCE model can be applied to individual patients in practice, based on specific patient characteristics. The figure shows the model predictions as well as the uncertainty at patient level. In practice, it is possible to calculate multiple survival probabilities for a single patient by selecting various initial treatments. However, one must take care in selecting the therapies as not all treatment may be relevant for the patient. Additionally, the survival predictions may have an inherent selection bias that needs to be considered. Patients in the NCR dataset that received no treatment probably had a worse performance status than patients that did receive treatment. This may result in an underestimation for the prediction of survival for best supportive care. Although this effect is partly corrected by other predictors in the model, there may still be bias in the model predictions.

Thus, these statistics and other model outcomes could be used to inform patients and aid the decision process by showing the relative change in survival for individual patients between treatments. To allow for implementation in clinical practice, however, a visual format is needed. For this purpose, we have created an interactive web-interface for SOURCE.[33] Although a nomogram is commonly used to this end, this presentation format is unsuitable for SOURCE as it contains interaction variables. The web-interface also contains functionality to highlight viable treatment options on the basis of patient characteristics. This will aid the selection of relevant treatments for patients in SOURCE. After extensive testing in clinical practice, this SOURCE web-interface will be made freely available for the oncological community.

In conclusion, the SOURCE prediction model for overall survival in metastatic oesophageal and gastric cancer was created based on a large nationwide cohort. SOURCE has both a fair discrimination and indicates a good accordance between predicted and observed survival. SOURCE can be used in clinical practice to give patients a personalised insight into their prognosis and thereby stimulate shared decision making.

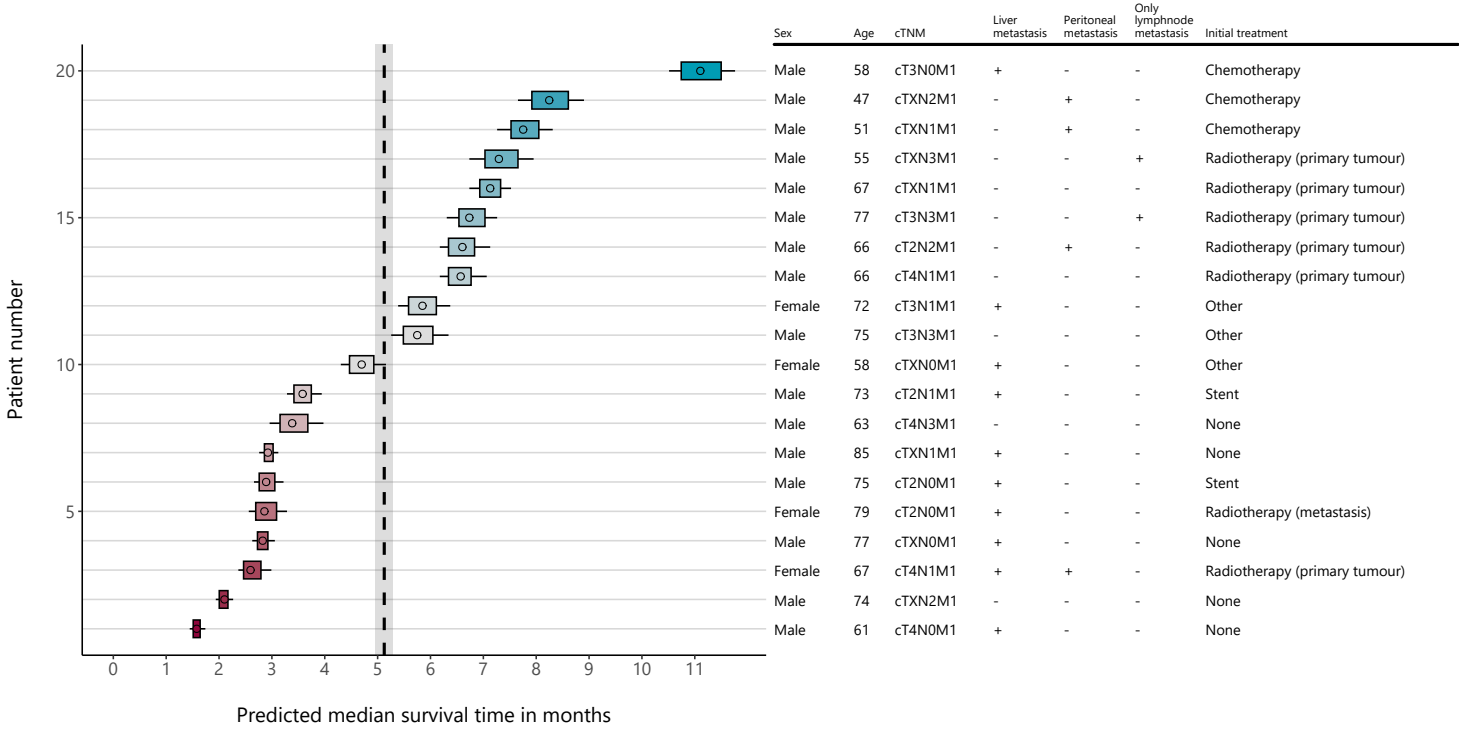


Figure 5. Predicted median survival times for metastatic oesophageal cancer.
The figure demonstrates the practical applicability of the SOURCE model in individual patients. The diagram is based on a random sample of 20 patients in the dataset. The SOURCE model predicts the median survival time with accompanying 50% confidence interval (bars) and 80% confidence intervals (lines). The dashed line indicates the observed median survival and confidence interval of all patients in the dataset. On the right, part of the patient characteristics are shown on which the predictions were based.

References

1. Enzinger, P. C. & Mayer, R. J. Esophageal Cancer. New England Journal of Medicine 349, 2241–2252 (2003).
2. Napier, K. J., Scheerer, M. & Misra, S. Esophageal Cancer: A Review of Epidemiology, Pathogenesis, Staging Workup and Treatment Modalities. World Journal of Gastrointestinal Oncology 6, 112–120 (2014).
3. Riihimäki, M., Hemminki, A., Sundquist, K., *et al.* Metastatic spread in patients with gastric cancer. Oncotarget 7, 52307–52316 (2016).
4. Wagner, A. D., Syn, N. L., Moehler, M., *et al.* Chemotherapy for advanced gastric cancer. Cochrane Database of Systematic Reviews 2017 (2017).
5. Lordick, F., Mariette, C., Haustermans, K., *et al.* Oesophageal cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Annals of Oncology 27, v50–v57 (2016).
6. Ajani, J. A., D'Amico, T. A., Almhanna, K., *et al.* Gastric Cancer, Version 3.2016; Clinical Practice Guidelines in Oncology. JNCCN Journal of the National Comprehensive Cancer Network 14, 1286–1312 (2016).
7. Al-Batran, S. E. & Ajani, J. A. Impact of chemotherapy on quality of life in patients with metastatic esophagogastric cancer. Cancer 116, 2511–2518 (2010).
8. Shah, M. A. Update on metastatic gastric and esophageal cancers. Journal of Clinical Oncology 33, 1760–1769 (2015).
9. Vickers, A. J. Prediction models in cancer care. CA: A Cancer Journal for Clinicians (2011).
10. Ravdin, P. M., Siminoff, L. A., Davis, G. J., *et al.* Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. Journal of Clinical Oncology 19, 980–991 (2001).
11. Van den Boorn, H. G., Engelhardt, E. G., van Kleef, J., *et al.* Prediction models for patients with esophageal or gastric cancer: A systematic review and meta-analysis. PLoS ONE 13 (2018).
12. Jung, H. A., Adenis, A., Lee, J., *et al.* Nomogram to predict treatment outcome of fluoropyrimidine/ platinum-based chemotherapy in metastatic esophageal squamous cell carcinoma. Cancer Research and Treatment 45, 285–294 (2013).
13. Shiozaki, H., Slack, R. S., Chen, H. C., *et al.* Metastatic Gastroesophageal Adenocarcinoma Patients Treated with Systemic Therapy Followed by Consolidative Local Therapy: A Nomogram Associated with Long-Term Survivors. Oncology (Switzerland) 91, 55–60 (2016).
14. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Annals of Internal Medicine 162, 55–63 (2015).
15. Ter Veer, E., van Rijssen, L. B., Besselink, M. G., *et al.* Consensus statement on mandatory measurements in pancreatic cancer trials (COMM-PACT) for systemic treatment of unresectable disease. The Lancet Oncology 19, e151–e160 (2018).
16. Ter Veer, E., van Kleef, J. J., Schokker, S., *et al.* Prognostic and predictive factors for overall survival (OS) in metastatic esophagogastric cancer (EGC): A meta-analysis. Journal of Clinical Oncology 35, 4048–4048 (2017).
17. Harrell Jr, F. E. rms: Regression Modeling Strategies. Tech. rep. (2018), 1–246.
18. R Core Team and others. R: A language and environment for statistical computing. 2013.
19. Fox, J. & Weisberg, S. Cox Proportional-Hazards Regression for Survival Data in R. Most 2008, 1–18 (2011).

20. Steyerberg, E. W. & Harrell, F. E. Prediction models need appropriate internal, internal-external, and external validation. Journal of Clinical Epidemiology 69, 245–247 (2016).
21. Van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. Journal of Statistical Software 45, 1–67 (2011).
22. Sobin, L., Gospodarowicz, M. & Wittekind, C. Classification of Malignant Tumours 6th edition (2002).
23. Edge, S., Byrd, D., Carducci, M. & Compton, C. AJCC cancer staging manual (2010).
24. Bozdogan, H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika 52, 345–370 (1987).
25. Obuchowski, N. A. Receiver operating characteristic curves and their use in radiology. Radiology 229, 3–8 (2003).
26. Steyerberg, E. W., Vickers, A. J., Cook, N. R., *et al.* Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology 21, 128–138 (2010).
27. Gerds, T. A. & Schumacher, M. Consistent estimation of the expected brier score in general survival models with right-censored event times. Biometrical Journal 48, 1029–1040 (2006).
28. Van der Heijden, G. J., T. Donders, A. R., Stijnen, T. & Moons, K. G. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. Journal of Clinical Epidemiology 59, 1102–1109 (2006).
29. Gensheimer, M. F., Henry, A. S., Wood, D. J., *et al.* Automated Survival Prediction in Metastatic Cancer Patients Using High-Dimensional Electronic Medical Record Data. Journal of the National Cancer Institute 111, 568–574 (2019).
30. Maltoni, M., Scarpi, E., Pittureri, C., *et al.* Prospective Comparison of Prognostic Scores in Palliative Care Cancer Populations. The Oncologist 17, 446–454 (2012).
31. Yates, J. W., Chalmer, B. & McKegney, F. P. Evaluation of patients with advanced cancer using the karnofsky performance status. Cancer 45, 2220–2224 (1980).
32. Henselmans, I., Van Laarhoven, H. W., Van der Vloodt, J., *et al.* Shared decision making about palliative chemotherapy: A qualitative observation of talk about patients' preferences. Palliative Medicine 31, 625–633 (2017).
33. Van den Boorn, H. G., van de Water, L. F., Hoxha, F., *et al.* Designing a user-friendly web-interface with prediction models for survival, health-related quality-of-life and toxicity for cancer patients (2018).

Supplement

	Predictor (% selected)
Round 1	
Included	Peritoneal metastasis (100%); surgery of primary tumour (100%); oesophagus vs. stomach tumour (100%); performance status (88%); liver metastasis (88%); disease status (locally advanced vs. metastasis) (88%); histology (Lauren) (75%); number of metastatic sites (75%); region/country (63%); HER status (63%); age (50%); disease status (unresectable vs. recurrent) (50%); histology (adenocarcinoma vs. squamous cell carcinoma) (50%); bilirubin (50%)
Rejected	White blood cell count (13%); Health-related quality of life (physical functioning) (13%); Health-related quality of life (role functioning) (13%); macroscopic tumour type (Borrmann) (13%); aspartate aminotransferase (13%); tumour size (13%); opt (13%); vascular endothelial growth factor a (13%); neuropilin (13%); MET expression (13%); pulmonary metastasis (0%); tuberous sclerosis (0%); sodium (0%); pylorus intact (0%); ERCC expression (0%); lymphnode metastasis (0%); peritoneal metastasis without drip infusion (0%); dihydropyrimidine dehydrogenase (0%); thymidine phosphorylase (0%); epidermal growth factor receptor (0%); lymphocytes (0%); neutrophils (0%); body surface area (0%); SPARC expression (0%); clinical N stage (0%)
Possible predictors	Alkaline phosphatase level (38%); GEJ vs. stomach tumour (38%); sex (38%); neutrophil to lymphocyte ratio (38%); peritoneal metastasis with ascites (38%); measurable disease (25%); Health-related quality of life (global health status) (25%); lactate dehydrogenase levels (25%); C-reactive protein (25%); ethnicity (25%); prior chemotherapy (25%); prior radiation (25%); weight loss (25%); Charlson comorbidity index (25%); VEGF (25%); visceral (lung or liver) metastasis (25%); KRAS wild type (vs. mutation) (25%)
Proposed predictors	Bone metastasis; oral intake; brain metastasis; tumour microsatellite instability status; body mass index; pain scale; EBV tumour status; virrhosis
Round 2	
Included	Peritoneal metastasis with ascites (63%); weight loss (50%); tumour microsatellite instability status (50%)
Rejected	GEJ vs. stomach tumour (38%); neutrophil to lymphocyte ratio (38%); bone metastasis (38%); brain metastasis (38%); EBV tumour status (38%); alkaline phosphatase level (25%); sex (25%); C-reactive protein (25%); prior radiation (25%); measurable disease (13%); Health-related quality of life (global health status) (13%); lactate dehydrogenase levels (13%); prior chemotherapy (13%); KRAS wild type (vs. mutation) (13%); oral intake (13%); body mass index (13%); ethnicity (0%); Charlson comorbidity index (0%); VEGF (0%); visceral (lung or liver) metastasis (0%); pain scale (0%); cirrhosis (0%)

Supplementary Table 1: Overview of included and excluded variables during the modified Delphi consensus.
Variables were included in the final selection if at least 50% of participants included the variables in their selection. In the first round, variables selected by less than 20% of participants were rejected, and the remaining variables were assessed in the second round. In the second round, all variables that were selected by at least 50% of participants were included in the final selection and all remaining variables were rejected.

Chapter 4

External validation of the Dutch SOURCE survival prediction model in Belgian metastatic oesophageal and gastric cancer patients

Jessy Joy van Kleef*, Héctor G. van den Boorn*, Rob H.A. Verhoeven, Katrijn Vanschoenbeek, Ameen Abu-Hanna, Aeilko H. Zwinderman, Mirjam A.G. Sprangers, Martijn G.H. van Oijen, Harlinde de Schutter and Hanneke W.M. van Laarhoven

Based on:

J.J. van Kleef*, H.G. van den Boorn*, R.H.A. Verhoeven, *et al.* External validation of the Dutch SOURCE survival prediction model in Belgian metastatic oesophageal and gastric cancer patients. Cancers 12 (2020).

* *These authors contributed equally*

Abstract

The SOURCE prediction model predicts individualised survival conditional on various treatments for patients with metastatic oesophageal or gastric cancer. The aim of this study was to validate SOURCE in an external cohort from the Belgian Cancer Registry.

Data of Belgian patients diagnosed with metastatic disease between 2004 and 2014 were extracted (N = 4,097). Model calibration and discrimination (c-indices) were determined. A total of 2,514 patients with oesophageal cancer and 1,583 patients with gastric cancer with a median survival of 7.7 and 5.4 months, respectively, were included. The oesophageal cancer model showed poor calibration (intercept: 0.30, slope: 0.42) with an absolute mean prediction error of 14.6%. The mean difference between predicted and observed survival was -2.6%. The concordance index (c-index) of the oesophageal model was 0.64. The gastric cancer model showed good calibration (intercept: 0.02, slope: 0.91) with an absolute mean prediction error of 2.5%. The mean difference between predicted and observed survival was 2.0%. The c-index of the gastric cancer model was 0.66.

The SOURCE gastric cancer model was well calibrated and had a similar performance in the Belgian cohort compared with the Dutch internal validation. However, the oesophageal cancer model had not. Our findings underscore the importance of evaluating the performance of prediction models in other populations.

Introduction

Oesophagogastric cancer has a dismal prognosis. Patients diagnosed with metastatic disease face a median overall survival (OS) time of three to five months with best supportive care (BSC).[1,2] Survival is dependent on various prognostic factors and treatment type.[3] Patients with a relatively good Eastern Cooperative Oncology Group Performance Status (PS) of 0–2, may be eligible for chemotherapy, targeted therapy, or even palliative surgery.[4,5] Brachytherapy, external radiotherapy, or stent placement may be deployed to relieve symptoms, such as dysphagia, and/or to reduce tumour growth.[6,7] Palliative treatments often have uncertain and limited benefit while the treatment burden can be high. Ideally, shared decision-making should be applied where patient preferences and values are taken into account during decision making.[8] Accurate and balanced information about treatment options tailored to the individual patient should be provided. However, oncologists were found to rarely discuss the potential pros and cons of palliative treatment and the BSC option.[9–12] This may, at least in part, be due to the complexity of predicting outcomes for individual patients.[13]

Prediction models can aid such individual risk estimation. Additionally, they can help quantify risks and benefits in an understandable manner to patients which allows them to more actively participate in the decision-making process.[14,15] Such prediction models will only live up to their potential if they have the required model performance qualities. A recent review investigated published risk prediction models regarding oesophagogastric cancer and concluded that model performance is often poorly described and external validation is limited.[16] In addition, no models in the metastatic setting were of sufficient quality for use in clinical practice. We therefore developed the SOURCE model (stimulating evidence-based, personalised and tailored information provision to improve decision-making after oesophageal-gastric cancer diagnosis).[17] The model makes OS predictions based on prognostic factors for metastatic oesophagogastric cancer patients. The SOURCE model was developed on a nationwide Dutch population-based cohort selected from the Netherlands Cancer Registry. Predictions regarding OS are conditional on various treatment types. Details on the input parameters, development and internal validation of the model were previously published.[17]

External validation is needed to investigate the performance of the original Dutch model and to justify its use for other populations. The Belgian population was selected, because the neighbouring countries have an extensive population-based national cancer registry. Therefore, the aim of this study was to validate the SOURCE model on an external population-based cohort selected from the Belgian Cancer Registry (BCR).

Methods

This manuscript was written in accordance with the TRIPOD statement.[18] The SOURCE model aims to stimulate evidence-based, personalised and tailored information provision to improve decision-making after oesophagogastric cancer diagnosis. The model predicts overall survival for patients with metastatic oesophageal or gastric carcinoma (cM_1), who did not die within 14 days after diagnosis. Patients with only distant metastases located in the head or neck region fall outside the target population of SOURCE. Input parameters of the models include: age, cT-category, cN-category, tumour differentiation grade, number of metastatic sites, distant lymph node metastasis only, intra-thoracic and intra-abdominal lymph node metastasis and initial treatment. The gastric cancer model also includes sex as an input parameter and the oesophageal cancer model also includes peritoneal, liver and head and neck metastases, morphology and topography. Input parameters were measured at diagnosis, before the start of treatment. SOURCE is integrated into a web-interface and will be made freely available after extensive assessments in clinical practice. Physicians can use the model together with patients during the clinical consultation. Since medical terminology is present in the web-interface, it is recommended that physicians discuss the results from the model with the patient, in a way that is tailored to the patient's level of understanding. It should be noted that SOURCE is developed to be a decision-aid to stimulate shared and informed decision-making. It should not and cannot replace the expertise and clinical judgement of physicians.

Data source

The BCR covers more than 95% of the Belgian cancer population.[19] Patient and tumour characteristics were collected from the standard cancer registration database, which relies on notifications from both the clinical (oncology care programmes) and pathological (laboratories for pathological anatomy) network. Data regarding treatment were derived from reimbursement claims of health insurance companies. A detailed description of the BCR data and data sources is given in the Supplementary Methods. The use of BCR data for scientific purposes is regulated by Belgian law, excluding the need for written informed consent for this study.

Patients

All patients diagnosed between 2004 and 2014 with a primary tumour in the oesophagus/gastroesophageal junction or stomach (ICD-10: C15.0–C16.9) and a cM_1 status were identified in the BCR. Analyses were restricted to patients with a Belgian residence at time of diagnosis. Inclusion and exclusion criteria were in accordance with the criteria used to develop the SOURCE model.[17] As this study took place entirely within the legal framework of the Belgian Cancer Registry, no ethical approval of concerned patients was needed. We more concretely refer to the privacy law of 08/12/1992 Chapter III Art 9 §2, 2e a) and 2e b) which refers to the Health Law of 2006.

Procedures

Treatment type was classified as for the original SOURCE model.[17] Input parameters for initial treatment were: BSC (registered as 'no treatment' or if no anti-cancer or symptom relief treatment was registered), radiotherapy (aimed at primary tumour or metastases), chemotherapy, chemoradiotherapy, chemotherapy plus short-term ($\leq$28 days) radiotherapy, resection (aimed at primary tumour or metastases), stent placement or other treatment (all other treatments not mentioned above, like targeted therapy only). Missing data regarding input parameters were handled using multiple imputation by chained equations.[20] Tumour staging was based on the 7th edition of the TNM staging system. However, patients diagnosed prior to 2010 were staged according to the TNM 6th edition. Conditional multiple imputation was used to align the definitions. This procedure has been described previously for SOURCE.[17] Conditional multiple imputation based on the original SOURCE dataset was also used to impute data regarding the target location (primary tumour or metastases) if patients underwent radiotherapy, since this level of detail was not given.

Statistical analyses

The primary endpoint was prediction of six-month overall survival. Overall survival was defined as the time between the date of diagnosis and death, or the date of last follow-up if a patient was censored. Differences in median survival between the development and validation cohort were assessed using Cox regression. To assess model performance, a concordance index (c-index) was calculated, as well as a calibration slope, intercept, absolute error and differences between predicted and observed survival outcomes. Model calibration was assessed by measuring the goodness-of-fit and is described by the agreement between predicted and observed outcomes.

In case of a perfect prediction, the calibration line has a slope of 1 and an intercept of 0 (x=y). A linear model was applied to assess the calibration slope and intercept of the model. The model was evaluated for the entire cohort and pre-defined patient subgroups based on the model's input parameters.[17] Mean differences between predicted and observed survival were calculated only for patient subgroups greater than 50 patients.

A c-index was calculated to assess the discriminatory ability of SOURCE. The c-index estimate is the probability that for a random pair of patients, the patient with the highest survival indeed has a higher predicted survival estimate than the other patient. The value 0.5 indicates that the model does not perform better than chance. A value of 1 indicates perfect discrimination. C-indices <0.7 were rated as poor, 0.7–0.79 as fair, 0.8–0.89 as good and 0.9–1 as excellent.[16] The SOURCE model was re-estimated using the input of the Belgian dataset with the method that was applied to create the original Dutch SOURCE model. Model performance for the re-estimated model was also assessed by means of c-indices, calibration slopes, intercepts and absolute prediction errors.

Data availability

The data that support the findings of our study are available in the Belgian Cancer Registry.

Results

Overall, 4,097 patients diagnosed between 2004 and 2014 registered by the BCR were included. Figure 1 depicts the selection process stratified by oesophageal and gastric cancer patients.

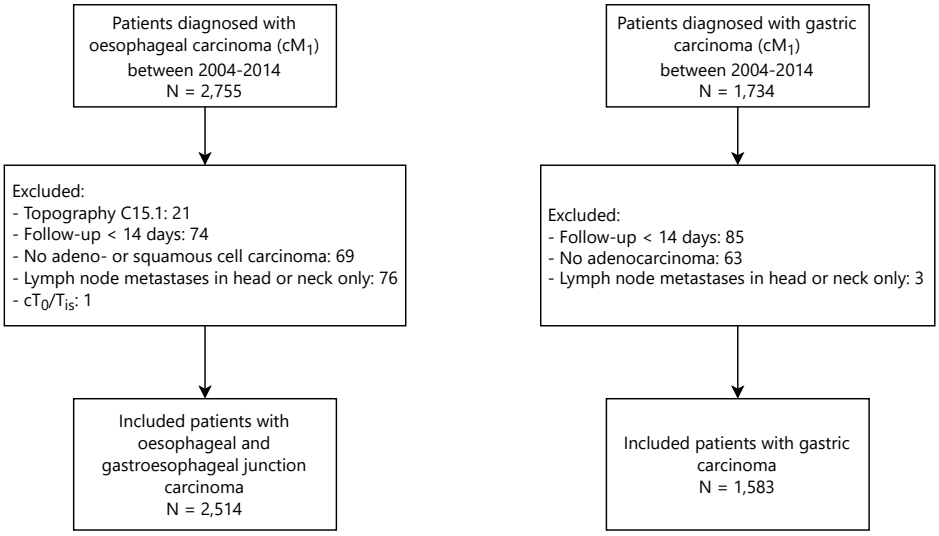


Figure 1: Flowchart showing inclusion of patients from the Belgian Cancer Registry in the study.

Oesophageal cancer patients

In total, 2,514 oesophageal cancer patients were analysed of whom 97.1% died during follow-up. Most patients were male (80.8%), had a PS of 1 (65.0%) and were diagnosed with adenocarcinoma (67.3%). The median observed OS was 7.7 months. An overview of patient, tumour and treatment characteristics is given in Table 1. Compared to the Dutch SOURCE population, the median OS time was higher for Belgian patients (7.7 vs. 5.1 months, $p < 0.0001$), see Table 1. cT_3 tumours were more frequently observed in Belgian patients (45.5% vs. 22.7%) whereas the Dutch population had a cT_X status in 49.9% of patients. Squamous cell carcinoma was compared to adenocarcinoma more frequently diagnosed in Belgium than in The Netherlands. Topography was not further specified in 33.1% of Belgian patients. Half of the Belgian patients were treated with chemotherapy, 10.6% received BSC and 5.8% received radiotherapy. Dutch patients received less treatment; 27.7% received chemotherapy, 26.6% BSC and 26% radiotherapy.

Patient subgroup	Belgian population, N (%)	Dutch SOURCE population, N (%)	Observed six-month OS (%)
All patients	2,514 (100)	8,010 (100)	58.2
Overall survival, median (IQR) in months	7.7 (3.2–15.3)	5.1 (2.2–10.1)	–
Sex			
Male	2,031 (80.8)	6,284 (78.5)	59.4
Female	483 (19.2)	1,726 (21.5)	53.4
Age			
Mean (SD)	65.7 (11.6)	66.8 (10.9)	–
<40	34 (1.4)	–	–
40–49	178 (7.1)	–	76.4
50–59	566 (22.5)	–	64.6
60–69	748 (29.8)	–	64.8
70–79	680 (27)	–	51.6
80–89	292 (11.6)	–	33.2
>90	16 (0.6)	–	–
Performance status			
Missing	248 (9.9)	–	–
0	258 (10.3)	–	71.5
1	1,634 (65)	–	61.5
2	274 (10.9)	–	40.2
3	83 (3.3)	–	20.8
4	17 (0.7)	–	–
cT category			
Missing	0 (0)	1 (0)	–
T_1	76 (3)	108 (1.3)	59.7
T_2	269 (10.7)	1,388 (17.3)	65.7
T_3	1,143 (45.5)	1,822 (22.7)	65.1
T_4	265 (10.5)	694 (8.7)	50.6
T_X	760 (30.2)	3,997 (49.9)	48.6
cN category			
Missing	172 (6.8)	1 (0)	–
N_0	694 (27.6)	2,127 (26.6)	48.7
N_1	706 (28.1)	2,502 (31.2)	63.2
N_2	670 (26.7)	2,391 (29.9)	62.7
N_3	272 (10.8)	989 (12.3)	59.5
Morphology			
Adenocarcinoma	1,692 (67.3)	6,321 (78.9)	60.1
Squamous cell carcinoma	790 (31.4)	1,423 (17.8)	55.3
Other	32 (1.3)	266 (3.3)	–
Topography primary tumour			
Cervical	15 (0.6)	44 (0.5)	–
Upper thoracic	94 (3.7)	205 (2.6)	55.3
Mid-thoracic	211 (8.4)	713 (8.9)	59.2
Lower thoracic	656 (26.1)	4,461 (55.7)	58.9
Overlapping lesion	6 (0.2)	315 (3.9)	–
Junction	701 (27.9)	2,112 (26.4)	62.3
Oesophagus NOS	831 (33.1)	160 (2)	53.4
Tumour differentiation grade			
Missing	373 (14.8)	3,472 (43.3)	–
G1	176 (7)	112 (1.4)	58.8
G2	822 (32.7)	1,464 (18.3)	59.5
G3	1,051 (41.8)	2,896 (36.2)	57.1
G4	92 (3.7)	66 (0.8)	57.8

Table 1 – *continued from previous page*

Patient subgroup	Belgian population, N (%)	Dutch SOURCE population, N (%)	Observed six-month OS (%)
Number of metastatic sites			
Missing	1,058 (42.1)	267 (3.3)	–
1	457 (18.2)	4,457 (55.6)	59.4
2	479 (19.1)	2,208 (27.6)	59.0
≥3	520 (20.7)	1,078 (13.5)	56.5
Lymph node metastases only			
Missing	1,058 (42.1)	267 (3.3)	–
No	1,320 (52.5)	6,532 (81.5)	56.6
Yes	136 (5.4)	1,211 (15.1)	70.5
Liver metastases			
Missing	1,058 (42.1)	267 (3.3)	–
No	655 (26.1)	3,699 (46.2)	61.5
Yes	801 (31.9)	4,044 (50.5)	56.1
Peritoneal metastases			
Missing	1,058 (42.1)	267 (3.3)	–
No	1,293 (51.4)	7,190 (89.8)	58.8
Yes	163 (6.5)	553 (6.9)	55.2
Head and neck LN metastases			
Missing	1,058 (42.1)	267 (3.3)	–
No	1,358 (54)	7,232 (90.3)	–
Yes	98 (3.9)	511 (6.4)	–
Intrathoracic LN metastases			
Missing	1,058 (42.1)	267 (3.3)	–
No	993 (39.5)	7,487 (93.5)	–
Yes	463 (18.4)	256 (3.2)	–
Intra-abdominal LN metastases			
Missing	1,058 (42.1)	267 (3.3)	–
No	990 (39.4)	6,218 (77.6)	–
Yes	466 (18.5)	1,525 (19)	–
First-line treatment			
None	266 (10.6)	2,131 (26.6)	19.9
Chemotherapy	1,247 (49.6)	2,216 (27.7)	71.8
Chemotherapy + short-term RT	277 (11)	317 (4)	74.0
Chemoradiotherapy	45 (1.8)	80 (1)	–
RT (primary tumour)	146 (5.8)	2,081 (26)	37.7
Resection (primary tumour)	60 (2.4)	0 (0)	–
RT (metastasis)	0 (0)	367 (4.6)	–
Resection (metastasis)	0 (0)	56 (0.7)	–
Stent	239 (9.5)	298 (3.7)	23.8
Other	234 (9.3)	464 (5.8)	51.7

Table 1: Observed six-month overall survival of the Belgian cohort and baseline characteristics of the development (Dutch) and validation (Belgian) oesophageal cancer cohort.

NOS: not otherwise specified, RT: radiotherapy; LN: lymph node.

SOURCE oesophageal cancer model validation

Model discrimination for the oesophageal cancer population amounted to a c-index of 0.64 (0.63–0.66), see Table 2. Model calibration at six months for the overall oesophageal cancer population corresponded to an intercept and calibration slope of 0.30 (0.28–0.31) and 0.42 (0.39–0.45), respectively. The mean difference between predicted and observed survival was -2.6% with a mean absolute prediction error of 14.6% (Table 2). The corresponding calibration plot (Figure 2) shows an underestimation of OS for patients with a predicted six-month OS of $\leq$46% with the most prominent deviations in the lowest tertile of the plot. Overestimation of six-month OS was present for patients with a relatively good prognosis, with larger deviations on the higher end of the scale (60–80%), see Figure 2. Figure 3 depicts mean differences between predicted and observed survival for various patient subgroups. The majority of patient subgroups (57%) showed larger differences between predicted and observed survival compared to the overall population (-2.6%).

Intercept	Slope	Absolute error (%)	Predicted-observed survival (%)	C-index
Oesophageal cancer model				
0.30 (0.28–0.31)	0.42 (0.39–0.45)	14.6 (14.5–14.7)	-2.6 (-4.3 – -1.0)	0.64 (0.63–0.66)
Gastric cancer model				
0.02 (0.02–0.02)	0.91 (0.90–0.91)	2.5 (2.5–2.5)	2.0 (1.8–2.2)	0.66 (0.64–0.68)

Table 2: Calibration and discriminative ability of the entire cohort at six months survival.

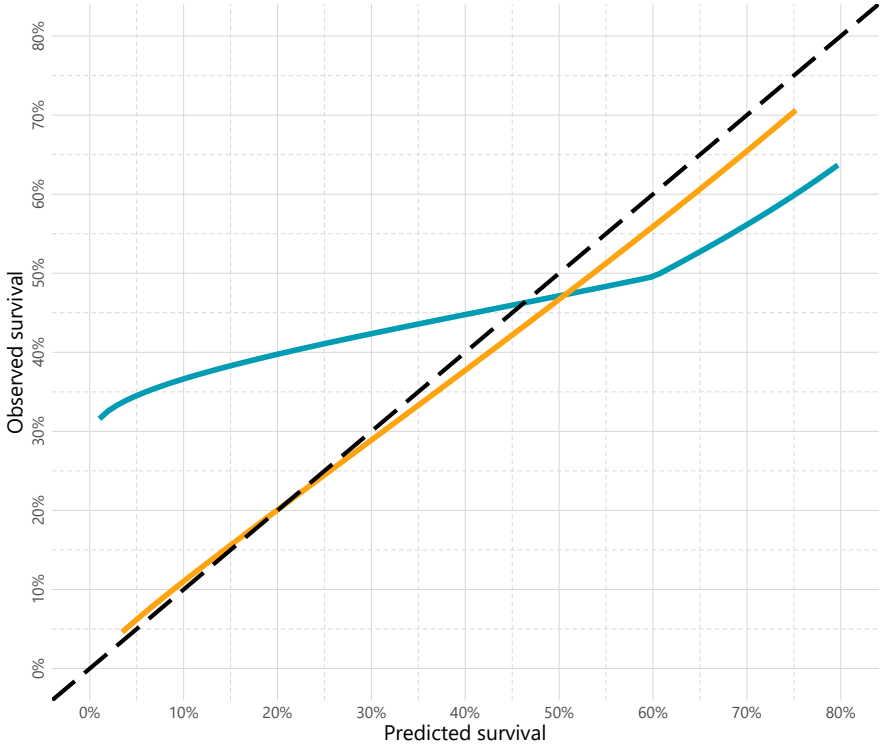


Figure 2: Calibration plot of predicted versus observed six-month overall survival.
The blue line indicates patients with oesophageal cancer and the orange line patients with gastric cancer.

Gastric cancer patients

In total, 1,583 patients with gastric cancer were analysed of whom 98.0% died during follow-up. Details of patient, tumour and treatment characteristics are given in Table 3. More than half of the patients were male (59.8%) and had a PS of 1 (59.6%). The median observed OS was 5.4 months. Compared to the original Dutch SOURCE population, the median OS time was longer for Belgian patients (5.4 vs. 3.9 months, $p < 0.0001$), see Table 3. The primary tumour location was not further specified in 60.1% of Belgian patients versus 8.4% of Dutch patients, and topography was assessed as an overlapping lesion in 0.5% of Belgian versus 34.5% of Dutch patients. Half (52.1%) of the Belgian patients were treated with chemotherapy versus 34.6% of Dutch patients, and 33.5% of Belgian patients received BSC versus 47.6% of Dutch patients. Detailed information regarding the location of metastases was missing in 658 (41.6%) Belgian patients.

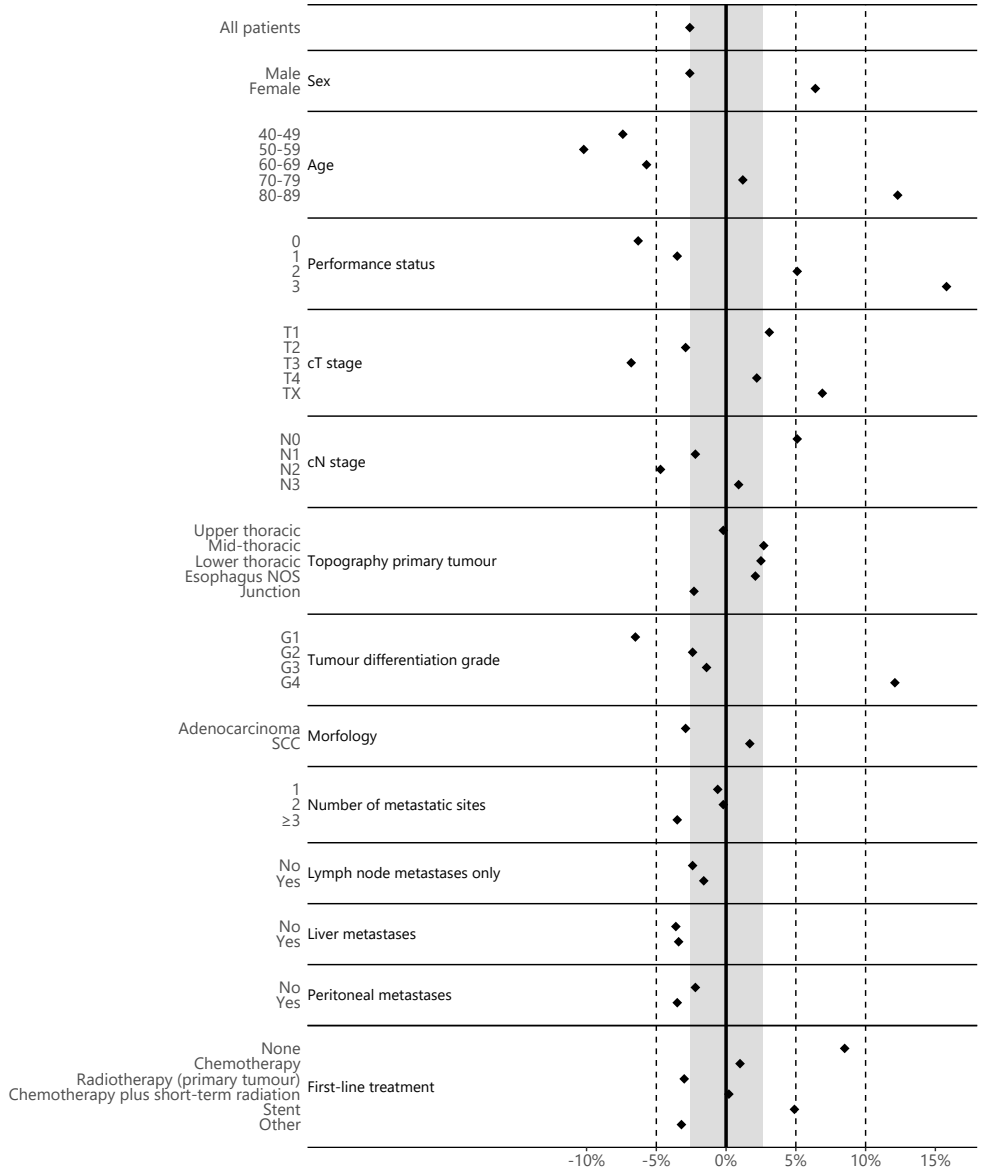


Figure 3: Mean differences between predicted and observed six-month overall survival for oesophageal cancer patients by patient subgroups.
Values $> 0\%$ indicate an overestimation and values $< 0\%$ indicate an underestimation in overall survival. The grey band represents the mean difference between predicted and observed six-month overall survival for the entire oesophageal cancer cohort. SCC: squamous cell carcinoma, NOS: not otherwise specified.

Patient subgroup	Belgian population, N (%)	Dutch SOURCE population, N (%)	Observed six-month OS (%)
All patients	1,583 (100)	4,763 (100)	46.7
Overall survival, median (IQR) in months	5.4 (2.1–11.9)	3.9 (1.7–8.4)	–
Sex			
Male	946 (59.8)	2,858 (60)	46.9
Female	637 (40.2)	1,905 (40)	46.3
Age			
Mean (SD)	70 (12.8)	68.6 (12.3)	–
<40	37 (2.3)	–	–
40–49	79 (5)	–	69.6
50–59	185 (11.7)	–	60.0
60–69	375 (23.7)	–	56.8
70–79	514 (32.5)	–	44.8
80–89	363 (22.9)	–	27.0
>90	30 (1.9)	–	–
Performance status			
Missing	131 (8.3)	–	–
0	143 (9)	–	67.3
1	944 (59.6)	–	50.4
2	255 (16.1)	–	34.4
3	79 (5)	–	17.1
4	31 (2)	–	–
cT category			
Missing	145 (9.2)	1 (0)	–
T_1	45 (2.8)	58 (1.2)	–
T_2	119 (7.5)	659 (13.8)	52.8
T_3	374 (23.6)	672 (14.1)	51.5
T_4	257 (16.2)	802 (16.8)	44.3
T_X	643 (40.6)	2,571 (54)	43.4
cN category			
Missing	141 (8.9)	0 (0)	–
N_0	767 (48.5)	2,366 (49.7)	44.6
N_1	259 (16.4)	1,012 (21.2)	48.1
N_2	336 (21.2)	1,264 (26.5)	49.4
N_3	80 (5.1)	121 (2.5)	50.9
Topography primary tumour			
Fundus	105 (6.6)	162 (3.4)	54.4
Corpus	158 (10)	954 (20)	47.5
Antrum Pylori	238 (15)	1,075 (22.6)	49.6
Pylorus	26 (1.6)	239 (5)	–
Lesser curvature NOS	63 (4)	181 (3.8)	50.8
Greater curvature NOS	34 (2.1)	106 (2.2)	–
Overlapping lesion	8 (0.5)	1,645 (34.5)	–
Stomach NOS	951 (60.1)	401 (8.4)	44.6
Tumour differentiation grade			
Missing	285 (18)	2,180 (45.8)	–
G1	101 (6.4)	42 (0.9)	48.5
G2	314 (19.8)	488 (10.2)	46.3
G3	847 (53.5)	2,028 (42.6)	46.7
G4	36 (2.3)	25 (0.5)	–
Number of metastatic sites			
Missing	658 (41.6)	168 (3.5)	–
1	355 (22.4)	3,099 (65.1)	46.0
2	273 (17.2)	1,067 (22.4)	45.1
≥3	297 (18.8)	429 (9)	48.9

Table 3 – *continued from previous page*

Patient subgroup	Belgian population, N (%)	Dutch SOURCE population, N (%)	Observed six-month OS (%)
Lymph node metastases only			
Missing	658 (41.6)	168 (3.5)	–
No	897 (56.7)	4,141 (86.9)	46.6
Yes	28 (1.8)	454 (9.5)	–
Liver metastases			
Missing	658 (41.6)	168 (3.5)	–
No	463 (29.2)	2,873 (60.3)	50.8
Yes	462 (29.2)	1,722 (36.2)	42.3
Peritoneal metastases			
Missing	658 (41.6)	168 (3.5)	–
No	459 (29)	2,735 (57.4)	45.5
Yes	466 (29.4)	1,860 (39.1)	48.0
Head and neck LN metastasis			
Missing	658 (41.6)	168 (3.5)	–
No	906 (57.2)	4,538 (95.3)	–
Yes	19 (1.2)	57 (1.2)	–
Intrathoracic LN metastasis			
Missing	658 (41.6)	168 (3.5)	–
No	839 (53)	4,419 (92.8)	–
Yes	86 (5.4)	176 (3.7)	–
Intra-abdominal LN metastasis			
Missing	658 (41.6)	168 (3.5)	–
No	642 (40.6)	3,973 (83.4)	–
Yes	283 (17.9)	622 (13.1)	–
First-line treatment			
None	531 (33.5)	2,266 (47.6)	16.8
Chemotherapy	825 (52.1)	1,648 (34.6)	64.5
Chemotherapy + short-term RT	28 (1.8)	52 (1.1)	–
Chemoradiotherapy	0 (0)	0 (0)	–
RT (primary tumour)	29 (1.8)	154 (3.2)	–
Resection (primary tumour)	111 (7)	247 (5.2)	63.1
RT (metastasis)	0 (0)	63 (1.3)	–
Resection (metastasis)	0 (0)	97 (2)	–
Stent	0 (0)	56 (1.2)	–
Other	59 (3.7)	180 (3.8)	45.8

Table 3: Observed six-month overall survival of the Belgian cohort and baseline characteristics of the development (Dutch) and validation (Belgian) gastric cancer cohort.

NOS: not otherwise specified, RT: radiotherapy; LN: lymph node.

SOURCE gastric cancer model validation

Model discrimination amounted to a c-index of 0.66 (0.64–0.68). Model calibration at six months for the overall gastric cancer population corresponded to an intercept and calibration slope of 0.02 (0.02–0.02) and 0.91 (0.90–0.91), respectively. The mean difference between predicted and observed survival was 2.0% with a mean absolute prediction error of 2.5% (Table 2). The corresponding calibration plot showed good calibration with no differences greater than 5% between predicted and observed survival along all prediction estimates, see Figure 2. Differences between predicted and observed survival were greatest in terms of overestimation for patients aged 80–89 (+6.1%), and with a PS score of 3 (+5.8%) and a cN_3 status (+5.5%). The majority of patient subgroups (59%) showed similar or smaller differences between predicted and observed OS compared to the overall cohort (-2.0%), see Figure 4.

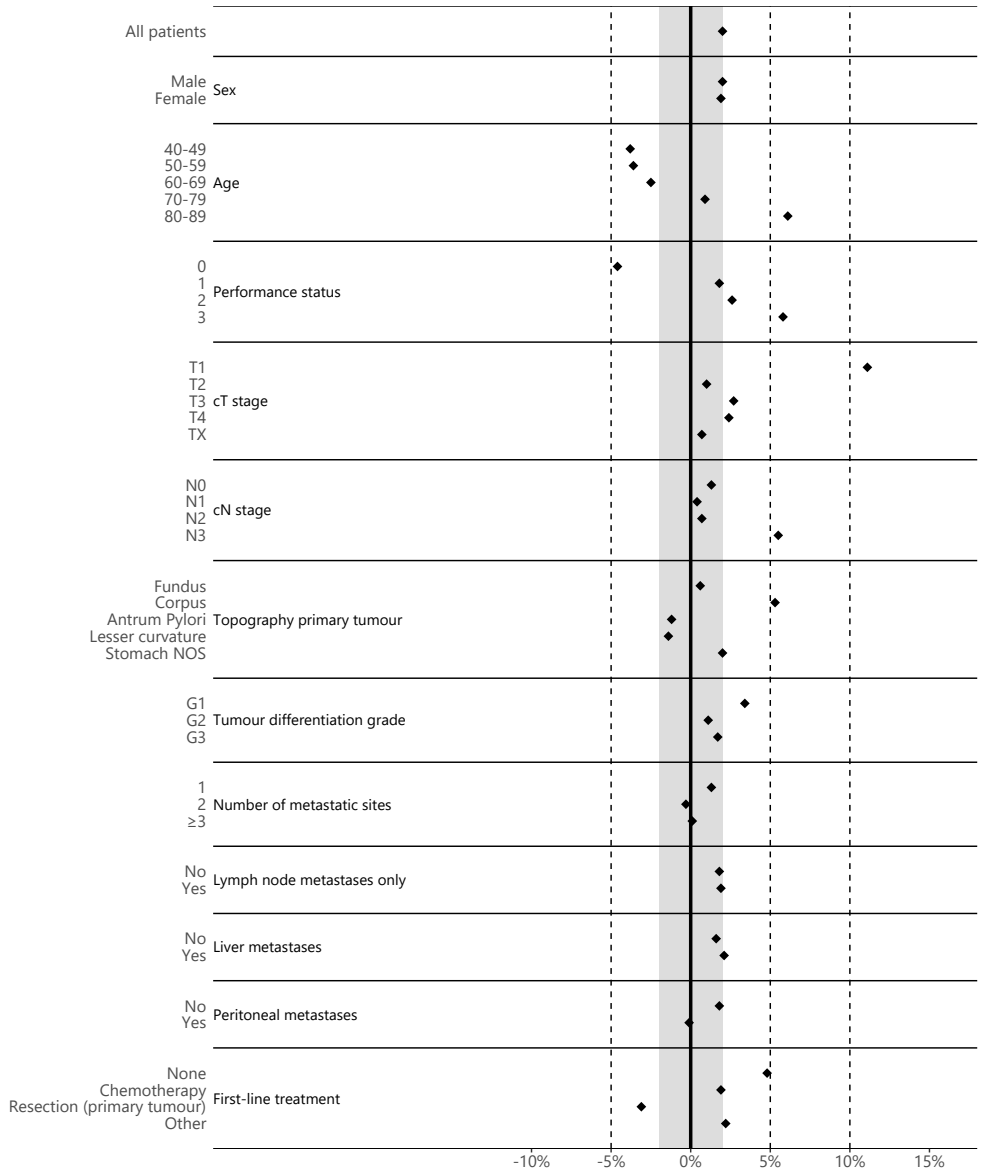


Figure 4: Mean differences between predicted and observed six-month overall survival for gastric cancer patients by patient subgroups.
Values $> 0\%$ indicate an overestimation and values $< 0\%$ indicate an underestimation in overall survival. The grey band represents the mean difference between predicted and observed six-month overall survival for the entire gastric cancer cohort. NOS: not otherwise specified.

Discussion

External validation of prediction models is essential for use in clinical practice.[21] This external validation study of the Dutch SOURCE model demonstrated that the oesophageal model had low transportability to the Belgian population, given its poor calibration and c-index. However, the gastric cancer model transported adequately. The original development report of SOURCE noted a calibration slope of one and an intercept of zero for both models during internal validation. C-indices were 0.71 and 0.68 for the oesophageal and gastric cancer model, respectively.[17] In this external validation study, we did not expect a superior performance compared to the internal validation, given the different nationality and healthcare settings. Our results showed that the oesophageal model performed poorer in the Belgian population regarding its calibration and a c-index of 0.64, but the performance of the gastric cancer model was close to the original internal validation with a good calibration and a c-index of 0.66.

Calibration of the gastric cancer model showed that differences between predicted and observed OS for the entire cohort were no greater than 5% along the calibration line, indicating a well calibrated model. Differences between predicted and observed OS were small ($<5\%$) for most patient subgroups. Older patients aged 80–90 had the largest difference ($+6.1\%$), which still was interpreted as fair by us. C-indices were relatively low according to our classification, indicating that the models had difficulties in making higher prediction estimates for patients who actually survived longer versus patients who had a shorter lifespan. Since the calibration was good for the gastric cancer population and the variation between prediction estimates was small, one might argue that the model had difficulties in ranking patients' OS.

The poor fit of the oesophageal cancer model might be explained by overfitting during model development. The oesophageal cancer model has more input parameters and interaction terms compared to the gastric cancer model. Such complex models with a high number of parameters might lead to good fit for the sample population—in this case the Dutch—but predictions might not generalise to new subjects outside the sample, such as the Belgians.[22] Missing data in the Belgian cohort might be another explanation for the poor fit. In this study, $>40\%$ of data regarding the location of metastases was missing and therefore multiple imputation was utilised to avoid selection bias. This, however, is always suboptimal in comparison to having observed values. The oesophageal model compared to the gastric model contains more input pa-

rameters regarding the location of metastases. Therefore, the oesophageal model validation was more subject to multiple imputation and thus uncertainty, which might explain the poorer fit.

Furthermore, adenocarcinoma and squamous cell carcinoma were combined into the same oesophageal cancer model, despite their differential biological features. Although the oesophageal cancer model contained histology as an input parameter, it is unclear to what extent this combination contributed to the poor model fit. Patient subgroup analysis showed that mean differences between predicted and observed survival for adenocarcinoma, squamous cell carcinoma, and the entire cohort were -2.9%, +1.7% and -2.6%, respectively. These mean differences did not substantially differ (see Figure 3). For the calibration and discrimination of the re-estimated model based on BCR data, see Supplementary Table 1 and Supplementary Figure 1.

Differences between development and validation datasets

Several differences in patient, tumour and treatment characteristics were observed between the Dutch and Belgian population. These include topography, cT-category and tumour differentiation grade, which might be due to missing data and/or differing cancer registration policies. In The Netherlands, data managers are centrally trained to interpret and register data in a standardised fashion. In Belgium, data collection is decentralised where clinical and pathological data is obtained by oncological care programmes and laboratories.[23] Albeit training of data managers and data cleaning is performed according to specific guidelines, differences in registration might thus be due to varying registration practices and/or interpretations.[24] In addition, BCR data regarding treatment types have been sufficiently validated. Data regarding the location of metastases were derived from Belgian hospital discharge data. This is the first study to use discharge records for this this purpose. It is, however, unknown to what extent this data might deviate from patients' medical records.

Taking patient selection into account, the proportion of patients with cM_1 tumours at diagnosis in the BCR was substantially lower compared to The Netherlands (22.1% vs. 40.1%). Additionally, the proportion of Belgian patients with a $cT_XN_XM_X$ status was considerably higher (28.6% vs. 1.9%; personal communication, 29 May 2019). So, one might argue that this Belgian $cT_XN_XM_X$ patient group is quite heterogeneous and that a portion of these patients had true cM_1 tumours at diagnosis. These patients, however, were not included in our analysis due to lack of detail in the clinical TNM classification. It might be the case that including these patients affects case mix and survival, which could lead to a dataset more similar to the Dutch.

When looking at the use of treatment modalities in the Belgian sample, the Belgians administered chemotherapy more frequently than the Dutch. Dutch oncologists more frequently offered BSC and radiotherapy, a more conservative approach that may explain the shorter median survival.[23–25] Lastly, SOURCE was developed to aid decision-making between BSC and (some form of) active treatment. During model development, 26.6% (N = 2,131) and 47.6% (N = 2,266) of Dutch oesophageal and gastric cancer patients received BSC (no treatment). Although this relatively large cohort could aid survival estimation on BSC, it should be pointed out that these estimates may have an inherent selection bias. Patients who received BSC most likely had worse PS scores or comorbidities compared to patients who did undergo treatment. Therefore, survival estimates for a relatively fit patient considering BSC may be underestimated. Although this effect could be partially corrected by other input parameters in the model, there may still be bias in the survival predictions.[17]

In conclusion, the SOURCE oesophageal model had low transportability to the Belgian population, but the gastric cancer model did transport adequately. Future studies should investigate the differences in diagnostics, treatment and survival between the populations, and the potential underlying causes. Model updating, in which newly available predictors can be incorporated to improve model performance, remains important. Furthermore, SOURCE should arm against overfitting by including fewer input parameters in future models. Lastly, for usage of the model in the Belgian clinical setting, model updating would be preferable in which ideally PS and more details regarding treatment could be incorporated.

References

1. Wagner, A. D., Syn, N. L., Moehler, M., *et al.* Chemotherapy for advanced gastric cancer. Cochrane Database of Systematic Reviews 2017 (2017).
2. Janmaat, V. T., Steyerberg, E. W., van der Gaast, A., *et al.* Palliative chemotherapy and targeted therapies for esophageal and gastroesophageal junction cancer. Cochrane Database of Systematic Reviews 2017 (2017).
3. Ter Veer, E., van Kleef, J. J., Schokker, S., *et al.* Prognostic and predictive factors for overall survival in metastatic oesophagogastric cancer: A systematic review and meta-analysis. European Journal of Cancer 103, 214–226 (2018).
4. Oncoline. Netherlands Comprehensive Cancer Organisation Richtlijn Oesofaguscarcinoom. 2015. `https://www.oncoline.nl/oesofaguscarcinoom` (2019).
5. Oncoline. Netherlands Comprehensive Cancer Organisation Richtlijn Maagcarcinoom. 2017. `https://www.oncoline.nl/maagcarcinoom` (2019).
6. Lordick, F., Mariette, C., Haustermans, K., *et al.* Oesophageal cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Annals of Oncology 27, v50–v57 (2016).
7. Smyth, E. C., Verheij, M., Allum, W., *et al.* Gastric cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Annals of Oncology 27, v38–v49 (2016).
8. Stiggelbout, A. M., Pieterse, A. H. & De Haes, J. C. Shared decision making: Concepts, evidence, and practice. Patient Education and Counseling 98, 1172–1179 (2015).
9. Koedoot, C. G., Oort, F. J., De Haan, R. J., *et al.* The content and amount of information given by medical oncologists when telling patients with advanced cancer what their treatment options are: Palliative chemotherapy and watchful-waiting. European Journal of Cancer 40, 225–235 (2004).
10. Rodriguez, K. L., Gambino, F. J., Butow, P., *et al.* Pushing up daisies: Implicit and explicit language in oncologist-patient communication about death. Supportive Care in Cancer 15, 153–161 (2007).
11. Gattellari, M., Voigt, K. J., Butow, P. N. & Tattersall, M. H. When the treatment goal is not cure: Are cancer patients equipped to make informed decisions? Journal of Clinical Oncology 20, 503–513 (2002).
12. Audrey, S., Abel, J., Blazeby, J. M., *et al.* What oncologists tell patients about survival benefits of palliative chemotherapy and implications for informed consent: Qualitative study. The BMJ 337, 492–496 (2008).
13. Glare, P., Virik, K., Jones, M., *et al.* A systematic review of physicians' survival predictions in terminally ill cancer patients. The BMJ 327, 195–198 (2003).
14. Belkora, J. K., Hutton, D. W., Moore, D. H. & Siminoff, L. A. Does use of the adjuvant! Model influence use of adjuvant therapy through better risk communication? JNCCN Journal of the National Comprehensive Cancer Network 9, 707–712 (2011).
15. Engelhardt, E. G., Garvelink, M. M., De Haes, J. C., *et al.* Predicting and communicating the risk of recurrence and death in women with early-stage breast cancer: A systematic review of risk prediction models. Journal of Clinical Oncology 32, 238–250 (2014).
16. Van den Boorn, H. G., Engelhardt, E. G., van Kleef, J., *et al.* Prediction models for patients with esophageal or gastric cancer: A systematic review and meta-analysis. PLoS ONE 13 (2018).
17. Van den Boorn, H. G., Abu-Hanna, A., Ter Veer, E., *et al.* SOURCE: A Registry-Based Prediction Model for Overall Survival in Patients with Metastatic Oesophageal or Gastric Cancer. Cancers 11 (2019).

18. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. BMC Medicine 13 (2015).
19. Belgian Cancer Registry. Cancer Burden in Belgium. Brussels, Belgium, 2015.
20. Van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. Journal of Statistical Software 45, 1–67 (2011).
21. Collins, G. S., De Groot, J. A., Dutton, S., *et al.* External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. BMC Medical Research Methodology 14 (2014).
22. Steyerberg, E. Clinical Prediction Models (Springer, New York, NY, 2009).
23. Claassen, Y. H., Bastiaannet, E., Hartgrink, H. H., *et al.* International comparison of treatment strategy and survival in metastatic gastric cancer. BJS open 3, 56–61 (2019).
24. Babaei, M., Balavarca, Y., Jansen, L., *et al.* Administration of adjuvant chemotherapy for stage II-III colon cancer patients: An European population-based study. International Journal of Cancer 142, 1480–1489 (2018).
25. Huang, L., Jansen, L., Balavarca, Y., *et al.* Stratified survival of resected and overall pancreatic cancer patients in Europe and the USA in the early twenty-first century: A large, international population-based study. BMC Medicine 16 (2018).

Supplement

Supplementary Methods

Data obtained from the BCR included age at diagnosis, sex, PS, year of diagnosis, topography, morphology, differentiation grade, (c/p)TNM stage, location of metastases, and information on systemic oncological, radiotherapeutic and surgical procedures up to six months after diagnosis. Patient and tumour characteristics were collected from the standard cancer registration database, which relies on notifications from both the clinical (oncology care programs) and pathological (laboratories for pathological anatomy) network.

Data regarding treatment were derived from reimbursement claims of health insurance companies as gathered by the Intermutualistic Agency. These data were linked to the BCR cancer registration data using the National Social Security Number (NSSN) as unique patient identifier, according to existing authorisations. All reimbursed diagnostic, therapeutic and pharmaceutical procedures from the in- and outpatient setting were available for a period from one year before until five years after diagnosis for each patient, limited to the end of 2017. Similarly, NSSN was used to retrieve information about vital status' from the Crossroads Bank of Social Security, updated until 1 July 2017.

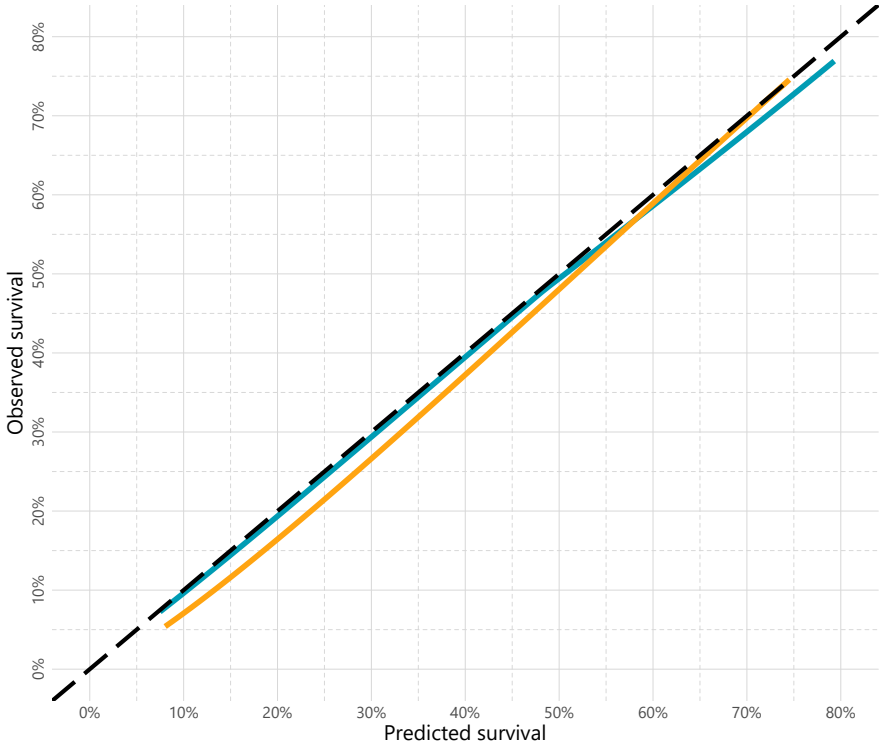
Data regarding the number of metastatic sites and their localisation was retrieved from the diagnostic codes (ICD-9-CM: 196, 197 and 198) within the time frame six months before until six months after incidence date as present in hospital discharge data available at BCR for this patient selection. These data were available for 57.9% of the patients. The use of BCR data for scientific purposes is regulated by Belgian law, excluding the need for written informed consent for this study.[17]

Patients who did not have adeno- or squamous cell carcinomas (SCC), who had a cT_0/T_{is} or tumour location C15.1 were excluded given SOURCE selection criteria. Patients with metastases only in lymph nodes of the head and neck area were also excluded. Additionally, patients who died within 14 days after diagnosis were excluded, given that prediction models are not likely to be applied for these patients.

Supplementary Results

Internal validation of the oesophageal cancer model at six months followup showed a c-index of 0.68 (0.66–0.69), an intercept of 0.00 (0.00–0.01), a slope of 0.97 (0.97–0.98) and an absolute mean error of 2.6% (Supplementary Figure 1). Mean differences between predicted and observed OS were +1.1%, see Supplementary Table 1. Internal validation of the gastric cancer model at six months follow-up showed a c-index of 0.68 (0.66–0.69), an intercept of

-0.04 (-0.05 – -0.04), a slope of 1.05 (1.05–1.06) and an absolute mean error of 2.3%, see Supplementary Figure 1. Mean differences between predicted and observed OS were also +2.3% (Supplementary Table 1).



Supplementary Figure 1: Calibration plot of predicted versus observed six-month overall survival.
The blue line indicates patients with oesophageal cancer and the orange line patients with gastric cancer. Results are shown from the re-estimated Belgian model.

Intercept	Slope	Absolute error (%)	Predicted-observed survival (%)	C-index
Oesophageal cancer model				
0.00 (-0.00–0.01)	0.97 (0.97–0.98)	2.6 (2.5–2.6)	1.1 (1.0 – 1.2)	0.68 (0.66–0.69)
Gastric cancer model				
-0.04 (-0.05 – -0.04)	1.05 (1.05–1.06)	2.3 (2.3–2.3)	2.3 (2.1–2.4)	0.68 (0.66–0.69)

Supplementary Table 1: Calibration and discriminative ability of the re-estimated model for the Belgium population at six months survival.

Chapter 5

SOURCE: Prediction models for overall survival in patients with metastatic and potentially curable oesophageal and gastric cancer

Héctor G. van den Boorn, Ameen Abu-Hanna, Nadia Haj Mohammad, Maarten C.C.M. Hulshof, Suzanne S. Gisbertz, Bastiaan R. Klarenbeek, Marije Slingerland, Laurens V. Beerepoot, Tom Rozema, Mirjam A.G. Sprangers, Rob H.A. Verhoeven, Martijn G.H. van Oijen, Aeilko H. Zwinderman and Hanneke W.M. van Laarhoven

Based on:

H.G. van den Boorn, A. Abu-Hanna, N. Haj Mohammad, *et al.* SOURCE: Prediction models for overall survival in patients with metastatic and potentially curable esophageal and gastric cancer. Journal of the National Comprehensive Cancer Network 19 (2021).

Abstract

Introduction

Personalised prediction of treatment outcomes can aid patients with cancer when deciding on treatment options. Existing prediction models for oesophagogastric cancer, however, have mostly been developed for survival prediction after surgery (i.e., when treatment has already been completed). Furthermore, prediction models for patients with metastatic cancer are scarce. The aim of this study was to develop prediction models of overall survival at diagnosis for patients with potentially curable and metastatic oesophageal and gastric cancer (the SOURCE study).

Methods

Data from 13,080 patients with oesophageal or gastric cancer diagnosed in 2015 through 2018 were retrieved from the prospective Netherlands Cancer Registry. Four Cox proportional hazards regression models were created for patients with potentially curable and metastatic oesophageal or gastric cancer. Predictors, including treatment type, were selected using the Akaike Information Criterion. The models were validated with temporal cross-validation on their c-index and calibration.

Results

The validated model's c-index was 0.78 for potentially curable gastric cancer and 0.80 for potentially curable oesophageal cancer. For the metastatic models, the c-indices were 0.72 and 0.73 for oesophageal and gastric cancer, respectively. The 95% confidence interval of the calibration intercepts and slopes contain the values 0 and 1, respectively.

Discussion

The SOURCE prediction models show fair to good c-indices and an overall good calibration. The models are the first in oesophageal and gastric cancer to predict survival at diagnosis for a variety of treatments. Future research is needed to demonstrate their value for shared decision-making in clinical practice.

Introduction

Oesophageal and gastric cancer are the eighth and fifth most common cancers worldwide, respectively.[1,2] Both types of cancer have a high mortality rate due in part to the high prevalence of distant metastases at diagnosis (20%–30% of patients with oesophageal cancer and 30%–40% of patients with gastric cancer).[2–4]

Multiple treatment options are available for patients with metastatic or potentially curable disease. Potentially curable gastric and oesophageal cancer can be treated with surgery with or without (neo)adjuvant chemo(radio)therapy. Potentially curable oesophageal cancer can also be treated with definitive chemoradiotherapy. Metastatic disease is treated mainly with systemic therapy but also with best supportive care.[5,6] However, even in the curative setting, outcome of these treatments is poor in oesophageal and gastric cancer, with 5-year survival rates <50%, whereas treatment-related morbidity is high.[7,8] Therefore, patient preferences and values should play a significant role in shared decision-making concerning treatment options.[9] When deciding on treatment, it is therefore vital that patients are provided accurate and preferably personalised information about the risks and benefits of the treatment trajectories, ideally based on prediction models.[10,11]

Several prediction models have been developed to predict risks and benefits in patients with oesophageal and gastric cancer.[12] However, these models are largely focused on the curative setting and predict survival after completion of a curative resection; they therefore cannot be used for clinical decision-making before the start of treatment, nor can they be used to compare different treatment options. Furthermore, prediction models for the metastatic setting are scarce.

We recently developed two models to predict survival in patients with metastatic oesophageal or gastric cancer based on tumour, patient, and treatment characteristics.[13] Although these models had good calibration and a fair c-index, some important information was not available at the time of development, such as HER2/neu status and WHO performance status, which would likely improve the models' performance.[14,15] The aim of this study (the SOURCE study) is to create two new models to predict survival in patients with potentially curable oesophageal or gastric cancer and to update our previously published models for patients with metastatic oesophageal and gastric cancer.

Methods

This article adheres to the TRIPOD guidelines.[16] According to the Central Committee on Research Involving Human Subjects, this type of study does not require approval from an ethics committee in The Netherlands. However, the study was approved by the Privacy Review Board of the Netherlands Cancer Registry (NCR; project code K17134).

Dataset

NCR data were used in the development and validation of the SOURCE prediction models.[17] This nationwide population-based registry is prospectively maintained. Since 2015, additional potential predictors were added to the database. We therefore decided to include only patients diagnosed in 2015 through 2018 (the year with the last available data in the NCR) with a primary oesophageal or gastric tumour. Patients with cM_1 disease were classified as having metastatic cancer, and patients with stage $cT_{1–4a,X}N_{0–3,X}M_0$ disease were classified as having potentially curable disease. This classification was also used in previous studies.[18,19] Patients with metastatic disease whose first metastasis was discovered at least four days after treatment initiation were classified as having cM_0 disease because they were diagnosed without any metastases.

Data from the NCR dataset were divided into four cohorts based on primary tumour location (oesophageal vs. gastric cancer) and cM stage (metastatic vs. potentially curable cancer). The primary tumour was classified as oesophageal cancer if the ICD-O topography code was C15.X or C16.0 (cardia) and as gastric cancer for C16.1–9.[20] Four prediction models were created based on these four datasets. The follow-up period lasted until January 2019 for a maximum of four-year follow-up for all patients.
Exclusion criteria included unknown vital status at the end of follow-up, unknown follow-up or survival of at most 14 days, primary cT_0 or in situ tumour, and unknown tumour histology. For patients with multiple primary tumours, duplicates were removed and only the earliest entry per patient was retained. Patients with metastatic disease who had distant metastases confined to lymph nodes of the head and neck area were excluded from the analyses because they could be treated with curative intent (Table 1).

	Oesophageal metastatic	Oesophageal potentially curable	Gastric metastatic	Gastric potentially curable
Initial selection	3,789	6,613	1,835	2,153
Unknown/short follow-up, vital status	182	104	134	63
cT_0 or cT_{is}	3	0	1	0
Unknown histology	63	61	52	26
Duplicates	3	43	2	4
Only head/neck metastases	267	–	17	–
Final selection	**3,271**	**6,405**	**1,629**	**2,060**

Table 1: Patient inclusion flowchart.
The four initial cohort sizes in the Netherlands Cancer Registry and the number of patients excluded are shown. The final selection was used in creating the SOURCE prediction models.

Development and validation of the prediction models

The methods for constructing the SOURCE prediction models were described in detail previously.[13] In short, the following procedures were followed. First, a preliminary predictor selection was made for each cohort. Predictors were selected if they were available for $\geq$50% in the dataset, had $<$50 levels (for categorical variables only), and did not have the same values for all patients (and would therefore have been noninformative). Performance status, body mass index (BMI), American Society of Anesthesiologists performance status classification, HER2/neu status, and laboratory results (hemoglobin, creatinine, lactate dehydrogenase, albumin levels) were also included in the preliminary predictor selection, in contrast to the previous study, because these variables became available for patients diagnosed as of 2015.[14,15] All predictors included in the SOURCE prediction models were determined at the time of diagnosis.

Next, a multivariate Cox proportional hazards regression model was created in each cohort, with overall survival as the outcome.[21] Overall survival was measured from diagnosis to death or censored at the date of last follow-up. In contrast to the previous study, the present models do not include interaction terms. It was found that the interaction terms did not increase model performance (data not shown), and the interaction terms were removed to avoid overfitting. Initially, all predictors from the preliminary selection were included in the model, and multiple imputation with 10 iterations via chained equations (multivariate imputation by chained equations [MICE]) was used to handle missing data.[22] A bidirectional predictor selection using the Akaike Information Criterion (AIC) was used to create a final predictor selection in each cohort.[23] From the resulting models, the c-index, calibration slope, intercept, and deviance were obtained. The c-index is a measure of discrimination and ranges from 0.5 (not able to discriminate survival outcome among indi-

viduals) to 1 (perfect discrimination).[24] The calibration refers to the concordance between predicted and observed survival. With a perfect calibration, the calibration intercept is 0, and the calibration slope is 1.[24] The calibration deviance refers to the mean absolute difference between predicted and observed survival.[25] All models were developed using the *rms* (regression modelling strategy) package in the RStudio environment with R version 3.6.1 (R Foundation for Statistical Computing).

To test the robustness of the models, an internal–external temporal cross-validation scheme was used.[26] Within this framework, the aforementioned model development was used to create models for patients diagnosed in earlier years, after which the model was evaluated based on patients in later years. This mimics the way in which models are evaluated when used in real life and reflects model performance behaviour in the face of potential population drift over time. This method allows the simulation of a true external temporal validation while using the entire available dataset.[26] This cross-validation is explained in more detail in our previous publication.[13] First, data from the earliest diagnosis year (2015) were used to create prediction models. The model's performance was then evaluated based on patients diagnosed in the subsequent year (2016). This process was then repeated for later diagnosis years; the training cohort included 2015 through 2016 and was validated based on the 2017 cohort, after which the model was trained on patients diagnosed in 2015 through 2017 and validated based on patients diagnosed in 2018. The performance statistics were pooled to obtain a cross-validated estimation of the model performance. To summarise, a Cox proportional hazards regression model was created for each cohort based on all available data, and a meta-analysis of cross-validated performance statistics was calculated to determine the model quality.

Results

Table 2 provides an overview of patient characteristics for the included cohort. Additional patient characteristics are provided in Supplementary Table 1. Kaplan-Meier curves for the four cohorts are provided in Supplementary Figure 1.

A complete overview of the parameters of the four SOURCE models is provided in Supplementary Table 2. This table shows the final predictor selection and the associated hazard ratios for each parameter in the multivariate Cox proportional hazards regression models. Table 3 shows an overview of the selected parameters in each prediction model.

	Oesophageal cancer N (%)		Gastric cancer N (%)	
	Metastatic	**Potentially curable**	**Metastatic**	**Potentially curable**
N (deaths)	3,271 (2,827)	6,405 (3,079)	1,629 (1,442)	2,060 (1,007)
OS, median (IQR), mo	5.3 (5.0–5.6)	22.7 (21.7–23.8)	4.3 (4.0–4.6)	22.8 (21.2–25.8)
Age mean (SD), y	67.40 (10.38)	69.59 (10.49)	68.95 (12.20)	72.19 (12.19)
Sex				
Male	2,568 (78.5)	4,643 (72.5)	987 (60.6)	1,238 (60.1)
Female	703 (21.5)	1,762 (27.5)	642 (39.4)	822 (39.9)
Weight, mean (SD), kg	77.26 (16.03)	77.80 (17.05)	73.69 (15.32)	73.96 (15.31)
BMI, mean (SD), kg/m^2	25.10 (4.46)	25.59 (4.74)	24.94 (4.57)	25.26 (4.51)
Missing	1,517 (46.4)	2,410 (37.6)	763 (46.8)	913 (44.3)
WHO performance status				
Missing	1,069 (32.7)	1,707 (26.7)	666 (40.9)	795 (38.6)
0	749 (22.9)	2,082 (32.5)	285 (17.5)	529 (25.7)
1	906 (27.7)	1,900 (29.7)	397 (24.4)	518 (25.1)
2	355 (10.9)	521 (8.1)	181 (11.1)	146 (7.1)
3+	192 (5.9)	195 (3.0)	100 (6.1)	72 (3.5)
cT stage				
cT_1	16 (0.5)	347 (5.4)	3 (0.2)	103 (5.0)
cT_2	1,186 (36.3)	1,891 (29.5)	453 (27.8)	773 (37.5)
cT_3	964 (29.5)	3,154 (49.2)	353 (21.7)	483 (23.4)
cT_4	240 (7.3)	109 (1.7)	281 (17.2)	91 (4.4)
cT_X	865 (26.4)	904 (14.1)	539 (33.1)	610 (29.6)
cN stage				
cN_0	596 (18.2)	3,210 (50.1)	701 (43.0)	1,457 (70.7)
cN_1	1,062 (32.5)	2,030 (31.7)	421 (25.8)	394 (19.1)
cN_2	1,237 (37.8)	1,003 (15.7)	430 (26.4)	181 (8.8)
cN_3	376 (11.5)	162 (2.5)	77 (4.7)	28 (1.4)
Initial treatment				
BSC	701 (21.4)	–	683 (41.9)	–
CRT	67 (2.0)	–	–	–
CT	999 (30.5)	–	587 (36.0)	–
Other	97 (3.0)	–	188 (11.5)	–
RT (metastasis)	219 (6.7)	–	–	–
RT (primary tumour)	954 (29.2)	–	108 (6.6)	–
Stent	234 (7.2)	–	63 (3.9)	–

Table 2 – *continued from previous page*

	Oesophageal cancer N (%)		Gastric cancer N (%)	
	Metastatic	Potentially curable	Metastatic	Potentially curable
Treatment				
BSC	–	529 (8.3)	–	405 (19.7)
CRT (high dose)	–	861 (13.4)	–	–
CRT (low dose)	–	557 (8.7)	–	–
CT	–	81 (1.3)	–	112 (5.4)
Endoscopic resection	–	362 (5.7)	–	61 (3.0)
NCRT and surgery	–	2,566 (40.1)	–	–
NCT and surgery	–	153 (2.4)	–	310 (15.0)
Other	–	227 (3.5)	–	79 (3.8)
Perioperative CT	–	132 (2.1)	–	471 (22.9)
RT	–	716 (11.2)	–	77 (3.7)
Resection	–	221 (3.5)	–	545 (26.5)

Table 2: Characteristics of included patients.
Patient characteristics are stratified by tumour location and metastatic versus potentially curable disease. cT stage, cN stage and differentiation grade are defined according to the TNM staging system of the 8th edition of the *AJCC Cancer Staging Manual*. The initial palliative treatment is the first treatment that a patient received after diagnosis. OS: overall survival; BMI: body mass index; IQR: interquartile range; SD: standard deviation; mo: months; y: years; BSC: best supportive care; NCRT: neoadjuvant chemoradiotherapy; CRT: chemoradiotherapy; CT: chemotherapy; NCT: neoadjuvant chemotherapy; RT: radiotherapy.

	Oesophageal cancer		Gastric cancer	
	Metastatic	Potentially curable	Metastatic	Potentially curable
Age	x	x	x	x
Sex	x			x
BMI*	x	x		x
WHO performance status*	x		x	
Albumine*	x	x	x	x
Hemoglobine*		x	x	
Lactate dehydrogenase*	x	x	x	x
Creatinine*	x			
cT stage	x	x	x	x
cN stage	x	x		x
Tumour topography		x		
Tumour topography				x
Morfology		x		
Differentiation grade	x	x	x	x
HER2 status*	x			
Only distance lymphnode metastasis	x			
Liver metastasis			x	
Peritoneal metastasis	x			
Number of metastatic sites	x		x	
Initial treatment	x		x	
Treatment		x		x

Table 3: Overview of selected parameters in each prediction model.
BMI: body mass index. *: parameters which were not available in the previous SOURCE models due to high missing rates.

The SOURCE models are also displayed graphically as nomograms in Supplementary Figures 2–5.[27] In a nomogram, the value of each predictor (e.g., the weight of the patient) is marked on its scale and then associated with a number of points that can be read from the top scale. The sum of all points can then be placed in the bottom scale, after which the survival estimate is determined. The nomograms provide survival estimates at six and 12 months for metastatic cancers and at one through four years for potentially curable cancers.

The performance statistics of all models are shown in Table 4. These results show an overall good calibration in all models. The 95% confidence intervals of the slopes and those of the intercepts include 1 and 0, respectively. The calibration deviance shows average prediction errors of 1% to 5%. The c-indices are 0.72 for metastatic cancers and are even higher for potentially curable cancers, with magnitudes of 0.78 and 0.80. Additional calibration plots are displayed in Supplementary Figure 6 and show the correspondence between predicted and observed survival per year cohort of validation.

	Metastatic		**Potentially curable**	
	Complete model[a]	**Validation**[b]	**Complete model**[a]	**Validation**[b]
Oesophageal cancer				
C-index	0.74 (0.72 – 0.75)	0.72 (0.71 – 0.74)	0.79 (0.78 – 0.80)	0.80 (0.75 – 0.84)
Calibration intercept	0.02 (0.01 – 0.02)	0.01 (-0.02 – 0.04)	0.05 (0.05 – 0.06)	-0.08 (-0.26 – 0.10)
Calibration slope	0.96 (0.96 – 0.97)	0.97 (0.90 – 1.04)	0.94 (0.93 – 0.94)	1.11 (0.88 – 1.35)
Calibration deviance	0.01 (0.01 – 0.01)	0.01 (0.01 – 0.02)	0.03 (0.02 – 0.03)	0.04 (0.03 – 0.07)
Gastric cancer				
C-index	0.73 (0.71 – 0.74)	0.72 (0.69 – 0.75)	0.78 (0.76 – 0.80)	0.78 (0.74 – 0.82)
Calibration intercept	-0.01 (-0.01 – -0.01)	-0.04 (-0.13 – 0.05)	0.01 (0.01 – 0.02)	-0.05 (-0.15 – 0.05)
Calibration slope	1.03 (1.03 – 1.03)	1.09 (0.96 – 1.22)	0.99 (0.99 – 1.00)	1.04 (0.90 – 1.18)
Calibration deviance	0.01 (0.01 – 0.01)	0.03 (0.02 – 0.04)	0.01 (0.01 – 0.01)	0.04 (0.03 – 0.06)

Table 4: SOURCE Model performance statistics.
[a]Complete model statistics refer to the statistics obtained when creating and testing the final prediction model using all available cohort data. [b]Validation statistics refer to the pooled values of all internal-external validation folds as described in the 'Methods' section.

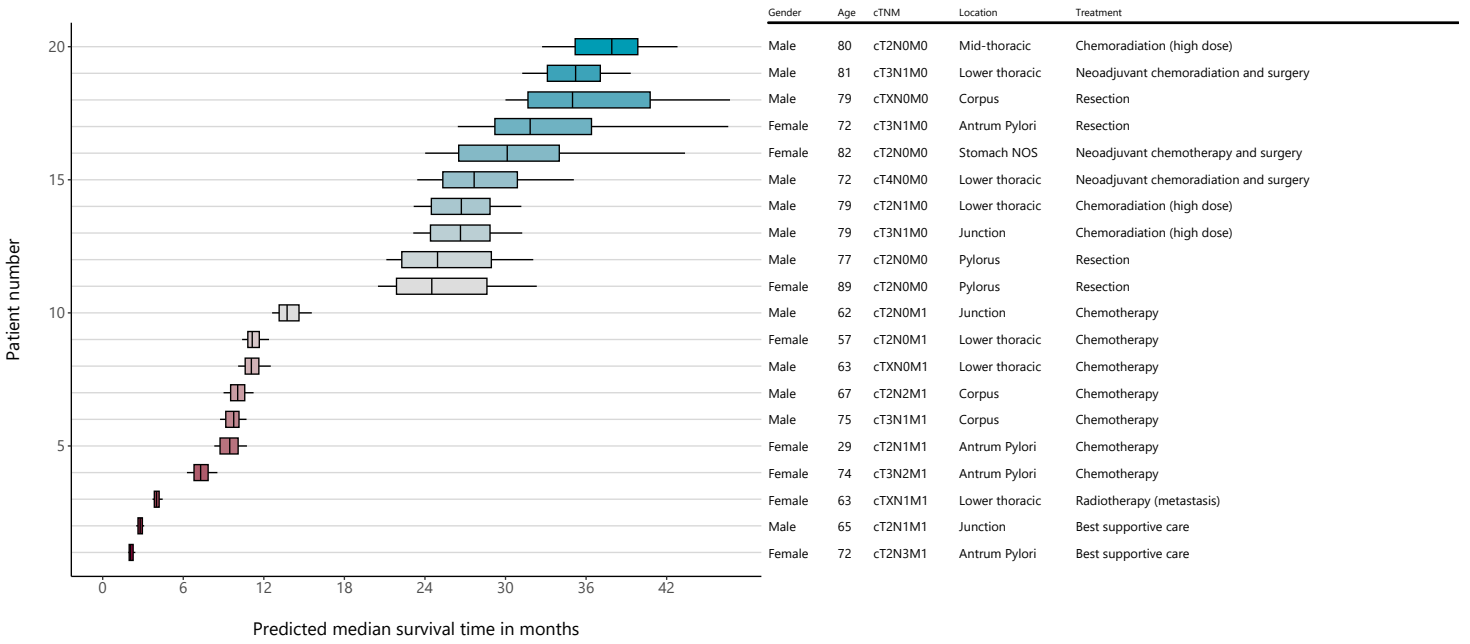


Figure 1: Individual predictions made by the SOURCE models.
The vertical line within each bar represents the predicted median survival for a random selection of patients. The bars show the 50% confidence interval, and the lines show the 80% confidence interval. The table on the right shows a selection of patient characteristics used for the predictions.

Discussion

The primary aim of this study was to create prediction models for overall survival in patients with potentially curable and metastatic oesophageal or gastric cancer. The SOURCE models are based on a large national cohort of patients diagnosed in recent years and form a complete set of models for use in upper gastrointestinal cancers. In contrast with other previously developed prediction models, the SOURCE models stand out due to their applicability to the full range of patients with curative and palliative oesophageal and gastric cancer and are to be used before the start of treatment.[28–30] Moreover, they are the first oesophageal and gastric cancer prediction models that include treatment as a predictor.

The robustness and generalisability of the models were considered during model development. The AIC method was used to automatically guide the predictor selection. Missing data were handled with multiple imputations (MICE). With this method, the prediction models are based on multiple datasets in which the missing values were imputed. The number of patients with at most two missing variables is 10,490 (78.5%). Because multiple imputations were made, the uncertainty of each individual imputation is taken into account.[31] This has the benefit of reducing bias compared with other methods, such as complete-case analysis.[31] To investigate the effect of overfitting, the models were also analysed with an internal–external temporal cross-validation. With this method, it is possible to simulate a temporal validation of the models that helps to examine how well the models might work with patient cohorts diagnosed in later years, provided they are more or less comparable.[26] This is especially relevant when developing models for clinical practice, because predictions will be made for patients diagnosed after the model has been developed.

The performance measures of the SOURCE models are similar for the complete model and for the internal–external cross-validation, indicating a lack of overfit. The c-indices of the potentially curable models are >0.75 (the average c-index of other prediction models for oesophageal and gastric cancer models), whereas the metastatic models had a c-index of 0.72 to 0.73, which can be considered fair.[12] There is also a good calibration slope and intercept for all models.

The presented metastatic models represent an update of our previously published models.[13] Model updating is an important part in the lifecycle of a prediction model.[32] The current models significantly differ from the previous models. First, the current models are developed based on more recent cohorts (2015–2018) than the previous models (2005–2015). In recent years, the NCR has extended its data collection to incorporate additional variables that could potentially be included as predictors and improve model performance. Indeed, WHO performance status and HER2/neu status are now included in the SOURCE models, as are BMI and albumin, hemoglobin, lactate dehydrogenase, and creatinine levels.[14,15] Second, parameter interaction terms were removed from the models; this had no significant effect on model performance, and further decreases the potential of overfitting. The resulting updated models show stable or even increased performance statistics, and the c-index of the gastric cancer model increased from 0.68 to 0.73. The model calibration demonstrates results in the updated models similar to those of the previous SOURCE models.

Some limitations of the SOURCE models should be mentioned. Patients were included as of 2015, implying a relatively short follow-up period, particularly for the cohorts with potentially curable disease. In this case, it was not possible to increase the follow-up to 5 years. In future models, a longer follow-up will be available, allowing predictions over a longer period of time for curative cohorts.

Another limitation is that information about treatment intent is not included in the NCR because it includes only the treatments patients actually received. For example, patients who intended to receive a neoadjuvant chemotherapy and surgery but did not advance to surgery because of clinical deterioration are classified as having received definitive chemotherapy. Predictions for definitive chemotherapy, for example, are therefore based on patients who intended to undergo definitive chemotherapy and on those who did not proceed to surgery after neoadjuvant treatment, which are clinical situations with likely different survival estimates. Furthermore, limited treatment details in the NCR led to broad treatment categories, as shown in Table 2. These limitations should be considered when using the SOURCE models.

In addition, these models are based solely on a Dutch population, which may impact the generalisability of this study. External validation should be performed to further determine the robustness of the SOURCE models and applicability to other populations of patients with oesophageal and gastric cancer.[26] For this undertaking, it is vital to take into consideration the comparability of cohorts with respect to, for example, tumour histology and primary tumour origin.[33]

The main strength of the SOURCE models lies in their clinical applicability. SOURCE forms a complete set of models that cover both potentially curable and metastatic oesophageal and gastric cancer. The predictors used in the models are readily available in standard clinical care and do not require additional testing. The inclusion of treatment as a model parameter makes it possible to compare the survival for various relevant treatment options, which can help with shared decision-making.[11] Figure 1 illustrates how the SOURCE models can be used to create predictions. The median predicted survival and confidence intervals are displayed for various patients with metastatic and potentially curable disease. It is also possible to compare the survival for various treatments, although one must be aware that not all treatments are relevant for each patient.

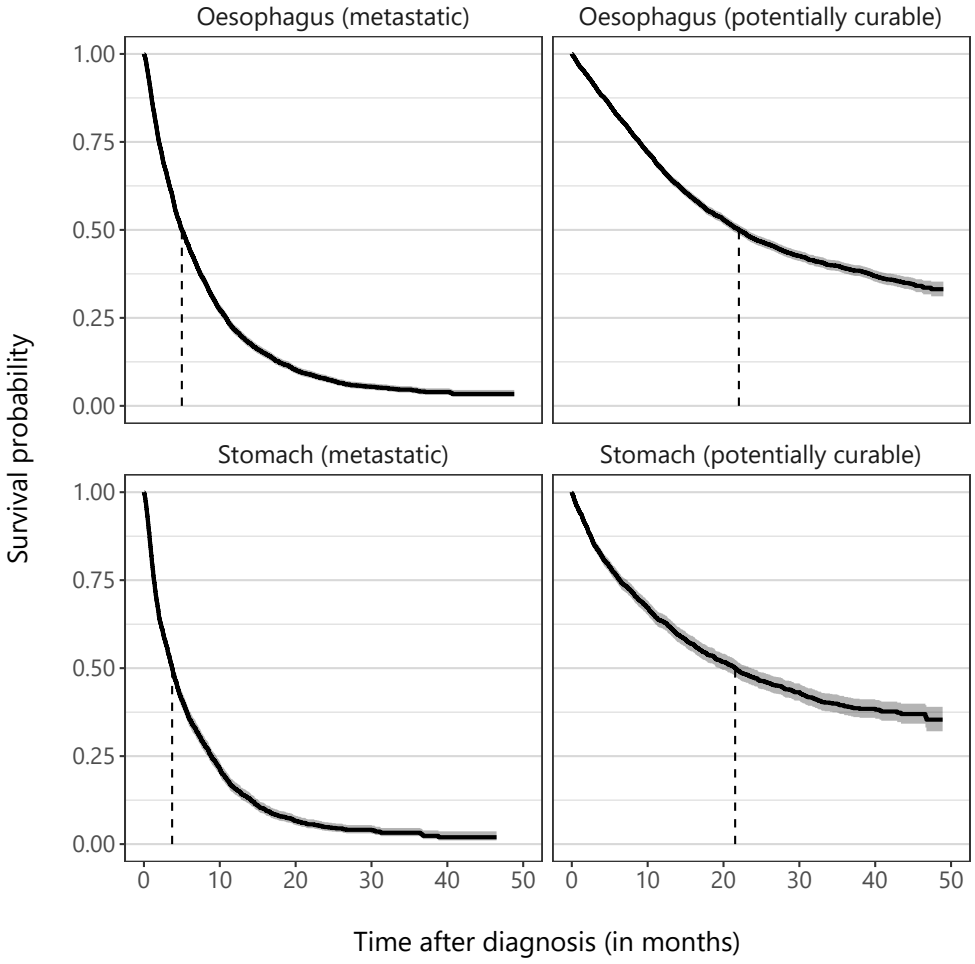
Currently, predictions can be made using the nomograms provided in Supplementary Figures 2–5. Although useful, these nomograms are not suitable for informing patients, and graphs or icon arrays should be used when informing patients about treatment outcomes.[34] The SOURCE models will be tested extensively in a clinical trial (ClinicalTrials.gov identifier: NCT04232735) to examine their effect on shared decision-making. The SOURCE models will become available through a web interface (https://source.amsterdamumc.org/) that is currently under development and the subject of a clinical trial, and they are therefore not accessible yet to the general public. This web interface will be used to facilitate the use of the prediction models and to display the predictions with user-friendly visualisations.

References

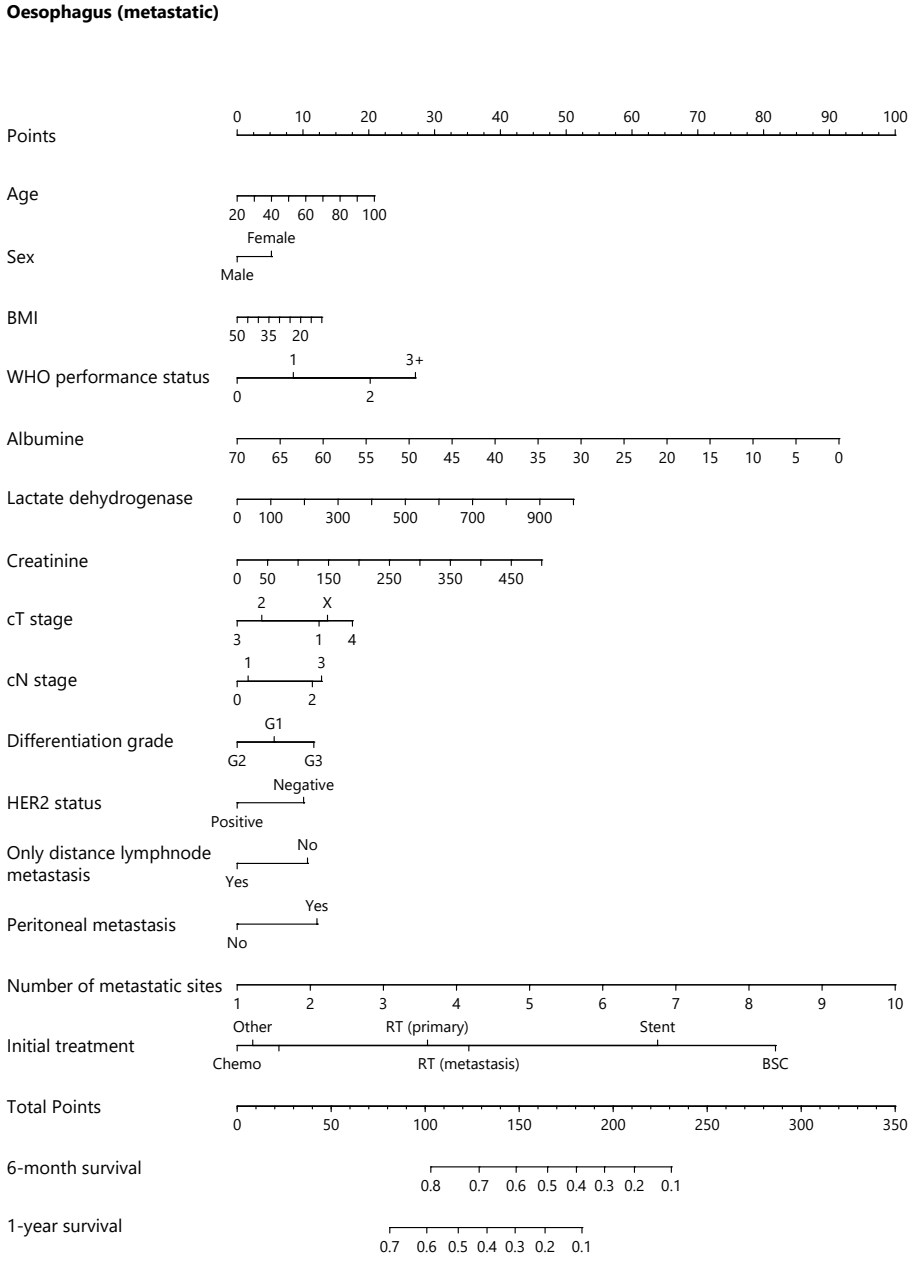
1. Arnold, M., Soerjomataram, I. & Ferlay, J. Global incidence of oesophageal cancer by histological subtype in 2012. Gut 64, 381–387 (2015).
2. Ferlay, J., Soerjomataram, I. & Dikshit, R. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 136, E359–386 (2015).
3. Napier, K. J., Scheerer, M. & Misra, S. Esophageal Cancer: A Review of Epidemiology, Pathogenesis, Staging Workup and Treatment Modalities. World Journal of Gastrointestinal Oncology 6, 112–120 (2014).
4. Riihimäki, M., Hemminki, A., Sundquist, K., *et al.* Metastatic spread in patients with gastric cancer. Oncotarget 7, 52307–52316 (2016).
5. Smyth, E. C., Verheij, M., Allum, W., *et al.* Gastric cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Annals of Oncology 27, v38–v49 (2016).
6. Lordick, F., Mariette, C., Haustermans, K., *et al.* Oesophageal cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Annals of Oncology 27, v50–v57 (2016).
7. Gavin, A. T., Francisci, S., Foschi, R., *et al.* Oesophageal cancer survival in Europe: A EUROCARE-4 study. Cancer Epidemiology 36, 505–512 (2012).
8. Van den Ende, T., Abe Nijenhuis, F. A., van den Boorn, H. G., *et al.* COMplot, A Graphical Presentation of Complication Profiles and Adverse Effects for the Curative Treatment of Gastric Cancer: A Systematic Review and Meta-Analysis. Frontiers in Oncology 9, 684 (2019).
9. Barry, M. J. & Edgman-Levitan, S. Shared Decision Making-The Pinnacle of Patient-Centered Care. The New England Journal of Medicine 366, 780–781 (2012).
10. Vickers, A. J. Prediction models in cancer care. CA: A Cancer Journal for Clinicians 61 (2011).
11. Henselmans, I., Van Laarhoven, H. W., Van der Vloodt, J., *et al.* Shared decision making about palliative chemotherapy: A qualitative observation of talk about patients' preferences. Palliative Medicine 31, 625–633 (2017).
12. Van den Boorn, H. G., Engelhardt, E. G., van Kleef, J., *et al.* Prediction models for patients with esophageal or gastric cancer: A systematic review and meta-analysis. PLoS ONE 13, e0192310 (2018).
13. Van den Boorn, H. G., Abu-Hanna, A., Ter Veer, E., *et al.* SOURCE: A registry-based prediction model for overall survival in patients with metastatic oesophageal or gastric cancer. Cancers 11 (2019).
14. Chan, D. S., Twine, C. P. & Lewis, W. G. Systematic Review and Meta-analysis of the Influence of HER2 Expression and Amplification in Operable Oesophageal Cancer. Journal of Gastrointestinal Surgery 16, 1821–1829 (2012).
15. Yates, J. W., Chalmer, B. & McKegney, F. P. Evaluation of patients with advanced cancer using the karnofsky performance status. Cancer 45, 2220–2224 (1980).
16. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Annals of Internal Medicine 162, 55–63 (2015).
17. Netherlands Comprehensive Cancer Organisation. IKNL and the NCR. 2020. `https://www.iknl.nl/en` (2020).
18. Van Putten, M., Verhoeven, R. H., Van Sandick, J. W., *et al.* Hospital of diagnosis and probability of having surgical treatment for resectable gastric cancer. British Journal of Surgery 103, 233–241 (2016).

19. Van der Sluis, P. C., van der Horst, S., May, A. M., *et al.* Robot-assisted Minimally Invasive Thoracolaparoscopic Esophagectomy Versus Open Transthoracic Esophagectomy for Resectable Esophageal Cancer: A Randomized Controlled Trial. Annals of surgery 269, 621–630 (2019).
20. Percy, C., Henson, D., Thomas, L. B. & Graepel, P. International classification of diseases for oncology (ICD-O). **7,** 402–403 (1977).
21. Fox, J. & Weisberg, S. Cox Proportional-Hazards Regression for Survival Data in R. Most 2008, 1–18 (2011).
22. Van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. Journal of Statistical Software 45, 1–67 (2011).
23. Bozdogan, H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika 52, 345–370 (1987).
24. Steyerberg, E. W., Vickers, A. J., Cook, N. R., *et al.* Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology 21, 128–138 (2010).
25. Gerds, T. A. & Schumacher, M. Consistent estimation of the expected brier score in general survival models with right-censored event times. Biometrical Journal 48, 1029–1040 (2006).
26. Steyerberg, E. W. & Harrell, F. E. Prediction models need appropriate internal, internal-external, and external validation. Journal of Clinical Epidemiology 69, 245–247 (2016).
27. Kattan, M. W. & Marasco, J. What Is a Real Nomogram? Seminars in Oncology 37, 23–26 (2010).
28. Hagens, E. R., Feenstra, M. L., Eshuis, W. J., *et al.* Conditional survival after neoadjuvant chemoradiotherapy and surgery for oesophageal cancer. British Journal of Surgery 107, 1053–1061 (2020).
29. Woo, Y., Son, T., Song, K., *et al.* A novel prediction model of prognosis after gastrectomy for gastric carcinoma. Annals of Surgery 264, 114–120 (2016).
30. Kattan, M. W., Karpeh, M. S., Mazumdar, M. & Brennan, M. F. Postoperative nomogram for disease-specific survival after an R0 resection for gastric carcinoma. Journal of Clinical Oncology 21, 3647–3650 (2003).
31. Sterne, J. A., White, I. R., Carlin, J. B., *et al.* Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. The BMJ (Online) 339, 157–160 (2009).
32. Steyerberg, E. W. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2009.
33. Van Kleef, J. J., van den Boorn, H. G., Verhoeven, R. H., *et al.* External validation of the dutch SOURCE survival prediction model in belgian metastatic oesophageal and gastric cancer patients. Cancers 12 (2020).
34. Zipkin, D. A., Umscheid, C. A. & Keating, N. L. Evidence-based risk communication: a systematic review. Ann Intern Med 161, 270–280 (2014).

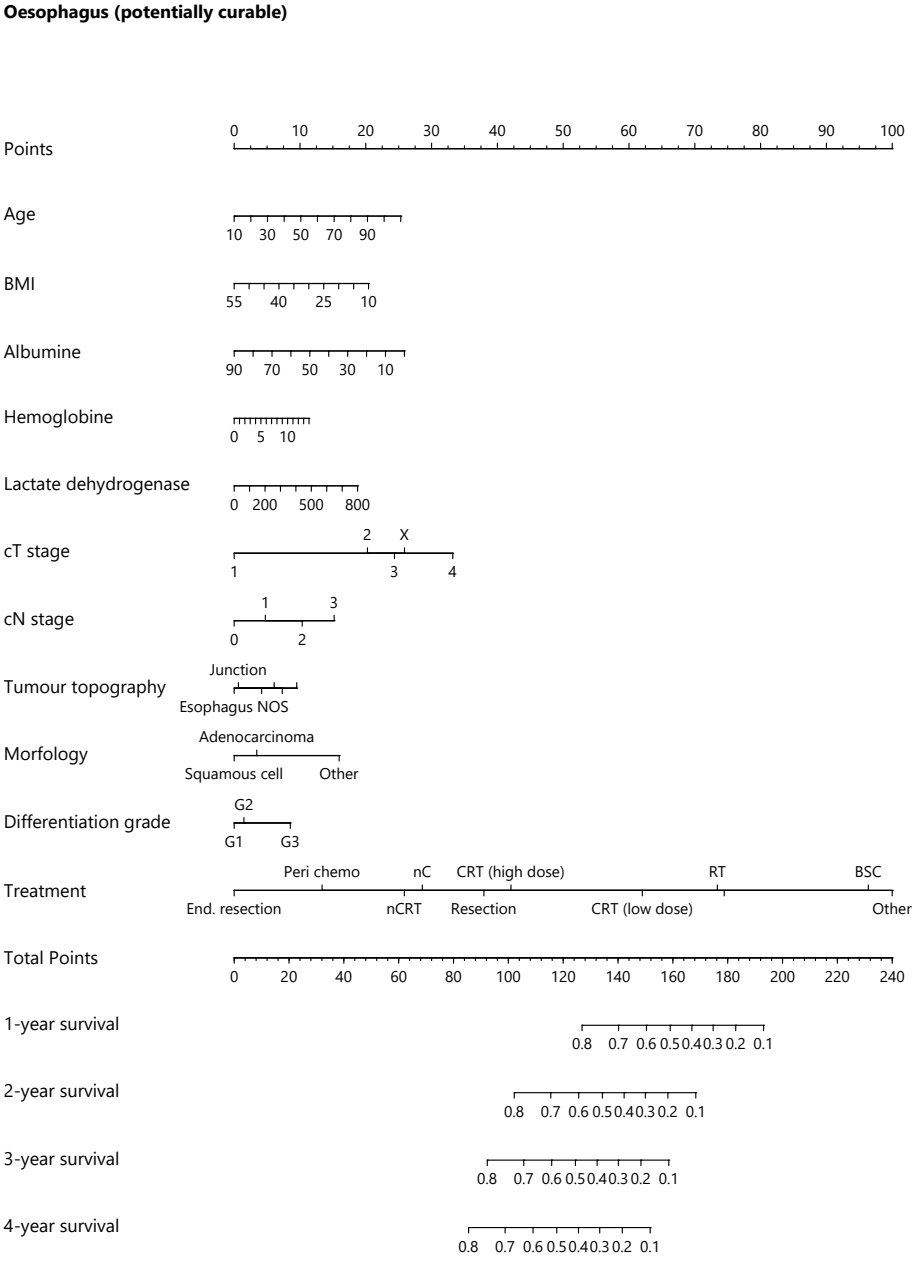
Supplement



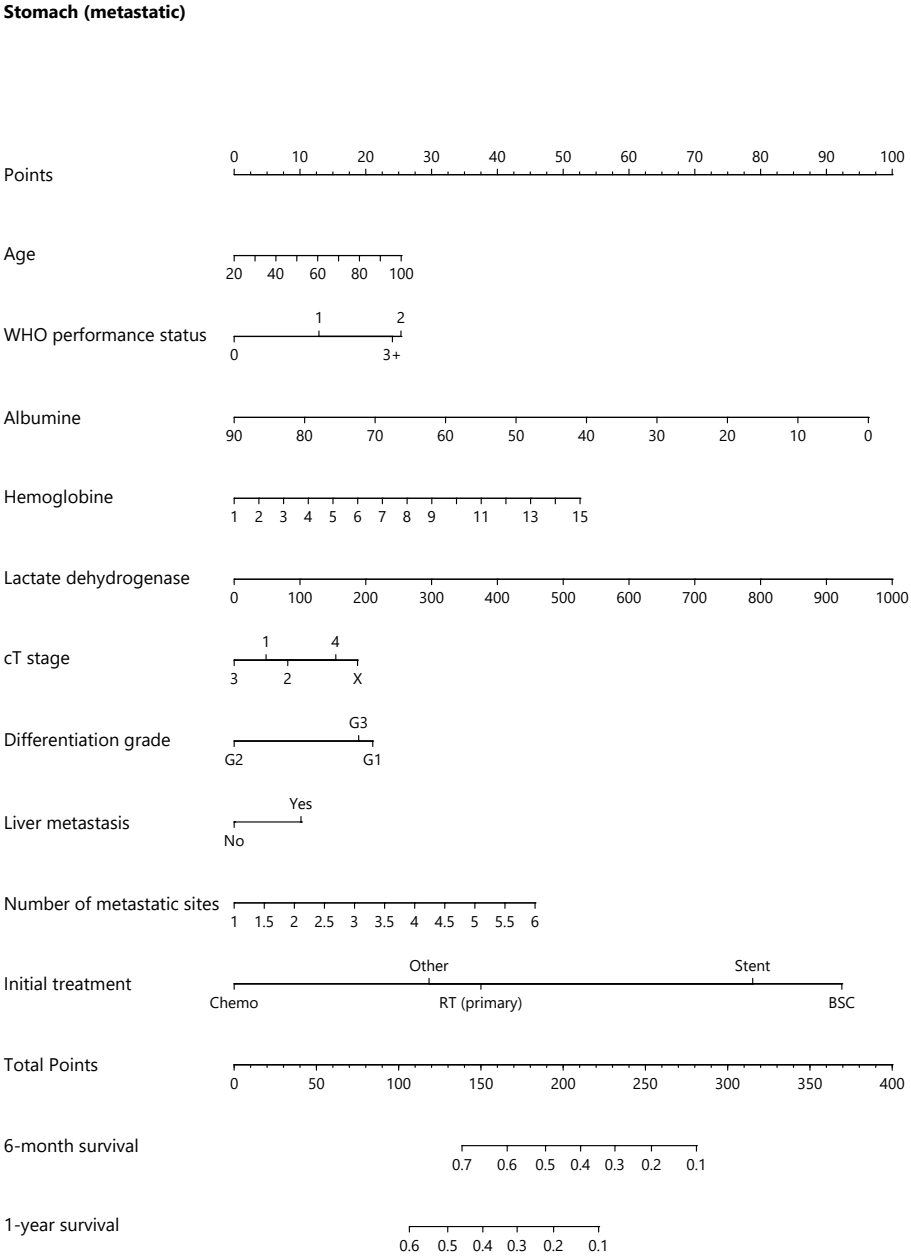
Supplementary Figure 1: Kaplan-Meier curves for the four cohorts.



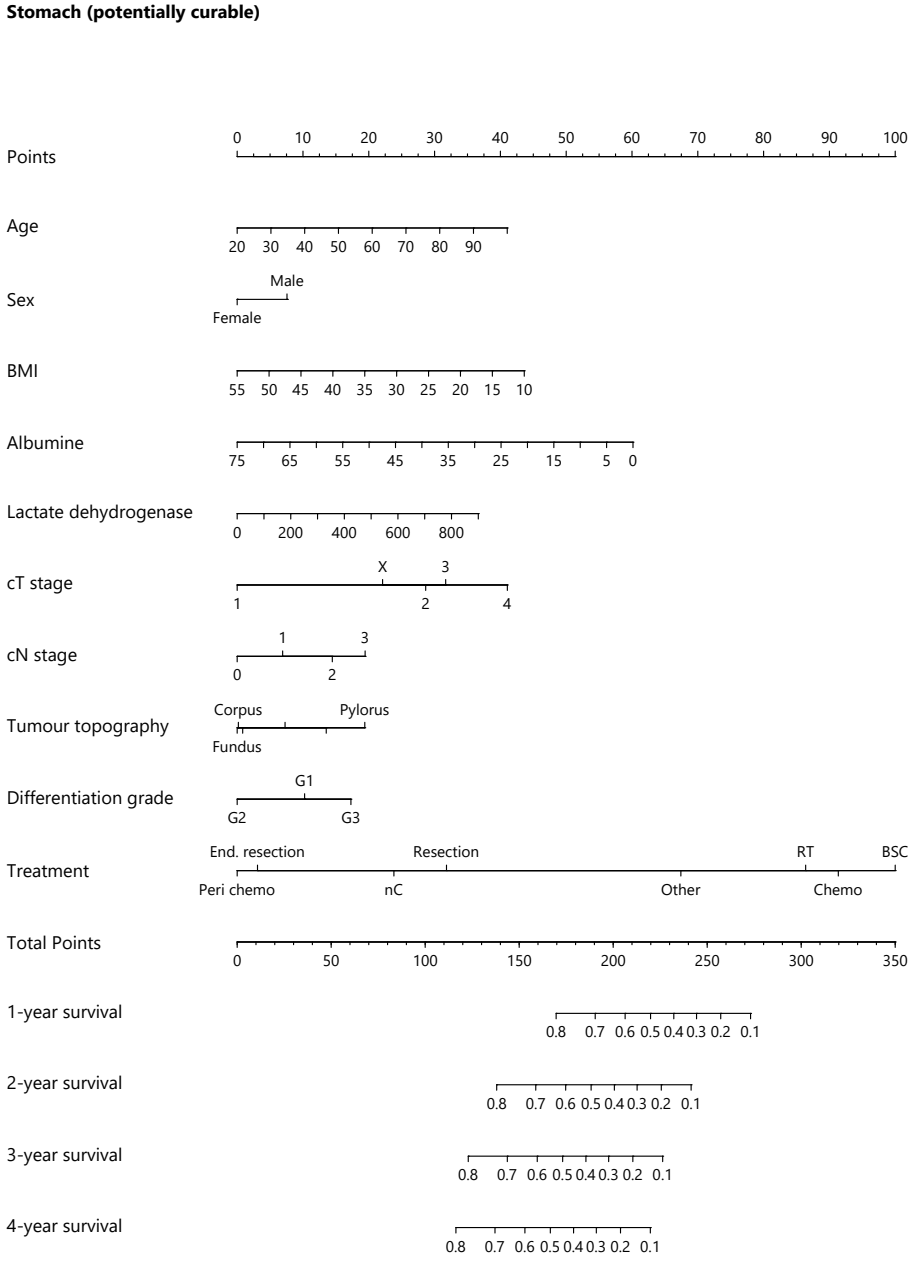
Supplementary Figure 2: Nomogram of model for oesophageal cancer: metastatic.



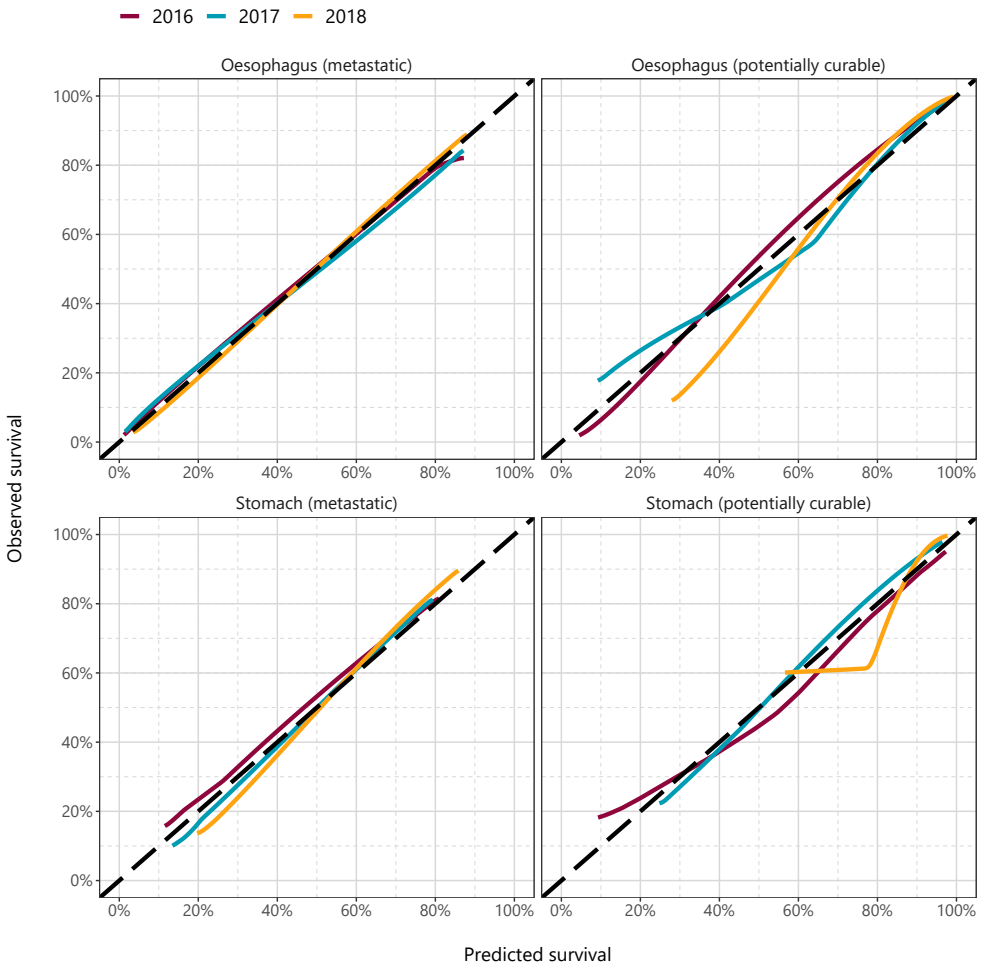
Supplementary Figure 3: Nomogram of model for oesophageal cancer: potentially curable.



Supplementary Figure 4: Nomogram of model for gastric cancer: metastatic.



Supplementary Figure 5: Nomogram of model for gastric cancer: potentially curable.



Supplementary Figure 6: Calibration plots for the four SOURCE models.

	Oesophageal cancer N (%)		Gastric cancer N (%)	
	Metastatic	Potentially curable	Metastatic	Potentially curable
ASA				
Missing	3,002 (91.8)	3,340 (52.1)	1,444 (88.6)	826 (40.1)
1	42 (1.3)	291 (4.5)	23 (1.4)	81 (3.9)
2	159 (4.9)	1,981 (30.9)	108 (6.6)	699 (33.9)
3	65 (2.0)	767 (12.0)	51 (3.1)	427 (20.7)
4	3 (0.09)	26 (0.4)	3 (0.2)	27 (1.3)
Albumine, mean (SD)	36.50 (6.44)	38.88 (5.78)	34.53 (6.99)	36.41 (6.48)
Missing	1,144 (35.0)	2,562 (40.0)	501 (30.8)	815 (39.6)
Hemoglobine, mean (SD)	8.11 (1.36)	8.41 (1.30)	7.24 (1.54)	7.38 (1.57)
Missing	140 (4.3)	499 (7.8)	62 (3.8)	128 (6.2)
LDH, (mean (SD))	256.83 (148.64)	190.76 (52.86)	240.31 (133.49)	195.95 (64.39)
Missing	756 (23.1)	1,566 (24.4)	296 (18.2)	510 (24.8)
Creatinine, mean (SD)	84.80 (32.06)	85.70 (32.06)	83.99 (35.89)	85.54 (33.80)
Missing	370 (11.3)	925 (14.4)	167 (10.3)	284 (13.8)
Tumour topography				
Lower thoracic	1,998 (61.1)	4,013 (62.7)	–	–
Upper thoracic	73 (2.2)	303 (4.7)	–	–
Mid-thoracic	281 (8.6)	884 (13.8)	–	–
Oesophagus NOS	127 (3.9)	191 (3.0)	–	–
Junction	715 (21.9)	922 (14.4)	–	–
Fundus	–	–	70 (4.3)	75 (3.6)
Corpus	–	–	417 (25.6)	509 (24.7)
Antrum Pylori	–	–	410 (25.2)	784 (38.1)
Pylorus	–	–	62 (3.8)	166 (8.1)
Stomach NOS	–	–	183 (11.2)	202 (9.8)
Overlapping lesion	77 (2.4)	92 (1.4)	487 (29.9)	324 (15.7)
Morfology				
Adenocarcinoma	2,720 (83.2)	4,669 (72.9)	1,611 (98.9)	2,042 (99.1)
Squamous cell	483 (14.8)	1,681 (26.2)	0 (0)	0 (0)
Other	68 (2.1)	55 (0.9)	18 (1.1)	18 (0.9)
Differentiation grade				
Missing	1,157 (35.4)	1,644 (25.7)	724 (44.4)	544 (26.4)
G1	98 (3.0)	342 (5.3)	19 (1.2)	44 (2.1)
G2	774 (23.7)	2,363 (36.9)	205 (12.6)	481 (23.3)
G3	1,242 (38.0)	2,056 (32.1)	681 (41.8)	991 (48.1)
HER2 status				
Missing	1,552 (47.4)	4,807 (75.1)	675 (41.4)	1,369 (66.5)
Negative	1,296 (39.6)	1,334 (20.8)	822 (50.5)	620 (30.1)
Positive	423 (12.9)	264 (4.1)	132 (8.1)	71 (3.4)
Headh/neck LNM				
No	3,023 (92.4)	–	1,599 (98.2)	–
Yes	248 (7.6)	–	30 (1.8)	–
Intra-thoracic LNM				
No	3,148 (96.2)	–	1,580 (97.0)	–
Yes	123 (3.8)	–	49 (3.0)	–
Intra-abdominal LNM				
No	2,659 (81.3)	–	1,409 (86.5)	–
Yes	612 (18.7)	–	220 (13.5)	–

Supplementary Table 1 – *continued from previous page*

	Oesophageal cancer N (%)		Gastric cancer N (%)	
	Metastatic	**Potentially curable**	**Metastatic**	**Potentially curable**
Only distance LNM				
No	2,788 (85.2)	–	1,489 (91.4)	–
Yes	483 (14.8)	–	140 (8.6)	–
Liver metastasis				
No	1,561 (47.7)	–	1,085 (66.6)	–
Yes	1,710 (52.3)	–	544 (33.4)	–
Peritoneal metastasis				
No	2,987 (91.3)	–	805 (49.4)	–
Yes	284 (8.7)	–	824 (50.6)	–
N metastatic sites, mean (SD)	1.76 (0.98)	–	1.54 (0.83)	–

Supplementary Table 1: Additional patient characteristics.
N: number; IQR: interquartile range; SD: standard deviation; ASA: American Society of Anesthesiologists; LDH: lactate dehydrogenase; LNM: lymph node metastasis; NOS: not otherwise specified.

	Oesophageal cancer		Gastric cancer	
	Metastatic	**Potentially curable**	**Metastatic**	**Potentially curable**
Age	1.005 (1.001 – 1.009)	1.009 (1.005 – 1.013)	1.004 (0.999 – 1.009)	1.012 (1.004 – 1.019)
Sex				
Male	1	–	–	1
Female	1.10 (1.00 – 1.21)	–	–	0.84 (0.74 – 0.96)
BMI	0.99 (0.99 – 1.00)	0.98 (0.98 – 0.99)	–	0.98 (0.97 – 0.99)
WHO PS				
0	1	–	1	–
1	1.17 (1.07 – 1.28)	–	1.18 (1.04 – 1.35)	–
2	1.45 (1.28 – 1.63)	–	1.39 (1.19 – 1.63)	–
3+	1.64 (1.41 – 1.90)	–	1.37 (1.14 – 1.64)	–
cT stage				
cT_1	1	1	1	1
cT_2	0.85 (0.50 – 1.44)	2.04 (1.55 – 2.69)	1.04 (0.33 – 3.29)	1.91 (1.20 – 3.05)
cT_3	0.80 (0.47 – 1.35)	2.36 (1.79 – 3.12)	0.94 (0.30 – 2.97)	2.04 (1.26 – 3.31)
cT_4	1.10 (0.64 – 1.88)	3.23 (2.26 – 4.61)	1.15 (0.36 – 3.63)	2.53 (1.48 – 4.33)
cT_X	1.02 (0.60 – 1.73)	2.49 (1.88 – 3.29)	1.20 (0.38 – 3.78)	1.65 (1.03 – 2.63)
cN stage				
cN_0	1	1		1
cN_1	1.03 (0.92 – 1.15)	1.18 (1.08 – 1.29)	–	1.17 (0.99 – 1.38)
cN_2	1.23 (1.10 – 1.37)	1.44 (1.29 – 1.60)	–	1.39 (1.11 – 1.73)
cN_3	1.26 (1.09 – 1.46)	1.71 (1.39 – 2.10)	–	1.55 (0.94 – 2.56)
Initial treatment				
BSC	1	–	1	–
CT	0.23 (0.20 – 0.25)	–	0.30 (0.26 – 0.34)	–
CRT	0.25 (0.19 – 0.34)	–	–	–
RT (primary)	0.38 (0.34 – 0.42)	–	0.49 (0.34 – 0.61)	–
RT (metastasis)	0.43 (0.36 – 0.51)	–	–	–
Stent	0.72 (0.62 – 0.84)	–	0.84 (0.64 – 1.10)	–
Other	0.24 (0.18 – 0.31)	–	0.44 (0.37 – 0.53)	–
Treatment				
BSC	–	1	–	1
NCRT and surgery	–	0.08 (0.07 – 0.10)	–	0.18 (0.14 – 0.23)
NCT and surgery	–	0.092 (0.07 – 0.13)	–	–
CRT (low dose)	–	0.30 (0.26 – 0.35)	–	–
CRT (high dose)	–	0.15 (0.13 – 0.17)	–	–
Resection	–	0.13 (0.10 – 0.17)	–	0.22 (0.18 – 0.26)
Perioperative CT	–	0.05 (0.04 – 0.08)	–	0.11 (0.08 – 0.14)
CT	–	0.46 (0.35 – 0.61)	–	0.82 (0.64 – 1.06)
RT	–	0.45 (0.39 – 0.51)	–	0.74 (0.56 – 0.96)
Endoscopic resection	–	0.03 (0.02 – 0.05)	–	0.11 (0.06 – 0.22)
Other	–	1.14 (0.96 – 1.34)	–	0.48 (0.35 – 0.66)
Albumine	0.98 (0.97 – 0.98)	0.99 (0.98 – 0.10)	0.99 (0.98 – 0.99)	0.98 (0.97 – 0.99)
LDH	1.00 (1.00 – 1.00)	1.03 (1.00 – 1.06)	1.05 (1.01 – 1.09)	1.00 (1.00 – 1.00)
Creatinine	1.00 (1.00 – 1.00)	1.00 (1.00 – 1.00)	1.00 (1.00 – 1.00)	–
Tumour topography				
Lower thoracic	–	1	–	–
Upper thoracic	–	1.12 (0.93 – 1.34)	–	–
Mid-thoracic	–	1.07 (0.95 – 1.21)	–	–
Oesophagus NOS	–	0.86 (0.67 – 1.07)	–	–
Junction	–	0.88 (0.79 – 0.99)	–	–
Antrum Pylori	–	–	–	1
Fundus	–	–	–	0.98 (0.69 – 1.40)

Supplementary Table 2 – *continued from previous page*

	Oesophageal cancer		Gastric cancer	
	Metastatic	**Potentially curable**	**Metastatic**	**Potentially curable**
Corpus	–	–	–	0.99 (0.83 – 1.17)
Pylorus	–	–	–	1.52 (1.19 – 1.94)
Stomach NOS	–	–	–	1.16 (0.93 – 1.44)
Overlapping lesion	–	1.21 (0.93 – 1.57)	–	1.33 (1.12 – 1.59)
Morfology				
Adenocarcinoma	–	1	–	–
Squamous cell	–	0.89 (0.80 – 0.98)	–	–
Other	–	1.56 (1.10 – 2.19)	–	–
Differentiation grade				
G1	1	1	1	1
G2	0.90 (0.78 – 1.05)	1.05 (0.90 – 1.24)	0.76 (0.61 – 0.94)	0.79 (0.60 – 1.05)
G3	1.12 (0.97 – 1.29)	1.35 (1.15 – 1.59)	0.97 (0.81 – 1.17)	1.17 (0.91 – 1.51)
HER2 status				
Negative	1	–	–	–
Positive	0.83 (0.76 – 0.91)	–	–	–
Only distance LNM				
No	1	–	–	–
Yes	0.82 (0.73 – 0.93)	–	–	–
Liver metastasis				
No	–	–	1	–
Yes	–	–	1.14 (1.01 – 1.29)	–
Peritoneal metastasis				
No	1	–	–	–
Yes	1.25 (1.09 – 1.43)	–	–	–
N metastatic sites	1.23 (1.18 – 1.28)	–	1.13 (1.06 – 1.20)	-

Supplementary Table 2: Prediction model parameters.

This table displays the models' hazard ratio (95% confidence interval) for the four SOURCE prediction models. N: number; BMI: body mass index; PS: performance status; ASA: American Society of Anesthesiologists; LDH: lactate dehydrogenase; LNM: lymph node metastasis; NOS: not otherwise specified; BSC: best supportive care; NCRT: neoadjuvant chemoradiotherapy; CRT: chemoradiotherapy; CT: chemotherapy; NCT: neoadjuvant chemotherapy; RT: radiotherapy.

Chapter 6

SOURCE-PANC: A prediction model for patients with metastatic pancreatic ductal adenocarcinoma based on nationwide population-based data

Héctor G. van den Boorn*, Willemieke P.M. Dijksterhuis*,
Lydia G.M. van der Geest, Judith de Vos-Geelen, Marc G. Besselink,
Johanna W. Wilmink, Martijn G.H. van Oijen and
Hanneke W.M. van Laarhoven

Based on:

H.G. van den Boorn*, W.P.M. Dijksterhuis*, L.G.M. van der Geest, *et al.* SOURCE-PANC: A prediction model for patients with metastatic pancreatic ductal adenocarcinoma based on nationwide population-based data. Journal of the National Comprehensive Cancer Network 19 (2021).

* *These authors contributed equally*

Abstract

Background

A prediction model for overall survival (OS) in metastatic pancreatic ductal adenocarcinoma (PDAC) including patient and treatment characteristics is currently not available, but it could be valuable for supporting clinicians in patient communication about expectations and prognosis. We aimed to develop a prediction model for OS in metastatic PDAC, called SOURCE-PANC, based on nationwide population-based data.

Methods

Data on patients diagnosed with synchronous metastatic PDAC in 2015 – 2018 were retrieved from the Netherlands Cancer Registry. A multivariate Cox regression model was created to predict OS for various treatment strategies. Available patient, tumour, and treatment characteristics were used to compose the model. Treatment strategies were categorised as systemic treatment (subdivided into FOLFIRINOX, gemcitabine/nab-paclitaxel, and gemcitabine monotherapy), biliary drainage, and best supportive care only. Validation was performed according to a temporal internal–external cross-validation scheme. The predictive quality was assessed with the c-index and calibration.

Results

Data for 4,739 patients were included in the model. Sixteen predictors were included: age, sex, performance status, laboratory values (albumin, bilirubin, CA19-9, lactate dehydrogenase), clinical tumour and nodal stage, tumour sublocation, presence of distant lymph node metastases, liver or peritoneal metastases, number of metastatic sites, and treatment strategy. The model demonstrated a c-index of 0.72 in the internal–external cross-validation and showed good calibration, with the intercept and slope 95% confidence intervals including the ideal values of 0 and 1, respectively.

Discussion

A population-based prediction model for OS was developed for patients with metastatic PDAC and showed good performance. The predictors that were included in the model comprised both baseline patient and tumour characteristics and type of treatment. SOURCE-PANC will be incorporated in an electronic decision support tool to support shared decision-making in clinical practice.

Introduction

In sharp contrast to mortality rates for virtually all other malignancies, those for pancreatic ductal adenocarcinoma (PDAC) have not declined since 2000.[1] Poor 5-year survival[2] is in part due to the fact that more than half of patients present with metastatic disease[3,4], and these patients usually have a prognosis of only a few months.[5,6] Predicting the exact survival time at diagnosis is challenging because of the heterogeneity of patients and tumours and differences in treatment of metastatic disease. Tools that can accurately predict survival while taking individual characteristics and treatments into account can be helpful for clinicians and patients when making treatment decisions.

Within the past decade, the emergence of prediction models in various cancer types has contributed to assessment of individually aligned prediction of prognosis and support of shared decision-making in clinical practice.[7] These models based on patient, tumour, and treatment characteristics aim to predict survival for individual patients. In PDAC, these models have been focused mainly on patients with localised disease.[8–14] The limited number of models that included patients with advanced disease were based on a small number of patients[15], developed for specific patient groups (e.g., patients after first-line systemic treatment[16] or treated with gemcitabine-based regimens[17–19]), based on patients included in trials only[19,20], limited to patients with locally advanced disease[17,21], or only described together with patients with localised diasease[22]. Moreover, they all did not take into account the different palliative systemic treatment options that are currently available for PDAC.[23–25]

Shared decision-making becomes increasingly important in clinical practice.[26] In patients with PDAC and metastases at initial diagnosis, median overall survival (OS) in the real-world setting (i.e., outside of clinical trials) ranges from 2.3 to 5.9 months in patients who receive best supportive care (BSC) only or systemic treatment, respectively.[6] Given the relatively marginal survival benefits, personalised patient information on treatment outcomes is of paramount importance. Multiple studies have shown that communication about prognosis is a challenge for physicians.[27] Prediction models can be helpful in supporting communication between physicians and patients regarding expectations and prognosis, and can enhance shared decision-making.[7]

The aim of this study was therefore to create a prediction model, called SOURCE-PANC (stimulating evidence-based, personalised, and tailored information provision to improve decision-making after pancreatic cancer diagnosis). The model was based on a large nationwide cohort using population-based data, with the goal of enabling prediction of OS from diagnosis in patients with metastatic PDAC undergoing palliative systemic treatment, biliary drainage, or BSC only.

Methods

Patient selection

Data from patients with a confirmed PDAC (C25 according to ICD-O-3[28]) or a nonconfirmed supposed adenocarcinoma and synchronous metastases ($T_{1–4,X}N_{0–2,X}M_1$) diagnosed between 2015 and 2018 were retrieved from the Netherlands Cancer Registry (NCR). The NCR is a population-based registry that includes all cancer diagnoses from the total Dutch population (>17 million inhabitants). Data on patient and tumour characteristics and treatment are extracted from medical records by trained registrars and include information about the patient (age, performance status, other cancer diagnosis, comorbidities), tumour (TNM stage, tumour biology, location of metastases), diagnosis (type of hospital, multidisciplinary consult, exploratory surgery), and treatment (systemic treatment, radiotherapy, palliative interventions, such as stents/drainages/bypasses, or BSC only). Data on vital status were obtained through annual linkage to the Dutch Personal Records Database and updated until February 1, 2020.

A total of 5,310 patients with metastatic PDAC were selected from the NCR. Patients aged <18 years or diagnosed at autopsy were not included. Patients who died within 14 days after diagnosis were excluded (N = 571); a prediction model is not applicable in these cases, because physicians will be able to predict the poor survival themselves in most cases and most probably would not consider starting a treatment other than BSC.[29] A total of 4,739 patients remained for development of the prediction model.

Treatment

Palliative treatment strategies were categorised as follows:

1. Systemic treatment: if patients received systemic therapy (N = 1,429);
2. Biliary drainage: if patients were not treated with systemic therapy but received a biliary stent or percutaneous biliary drainage (N = 722);
3. BSC: if patients did not receive systemic treatment or biliary drainage (N = 2,588).

Subsequently, systemic therapy regimens were subdivided as follows:

1. Fluoropyrimidine, platinum, and irinotecan (e.g., FOLFIRINOX [5-FU/leucovorin/oxaliplatin/irinotecan]);
2. Regimens with gemcitabine and nab-paclitaxel;
3. Gemcitabine monotherapy;
4. Other regimens.

Predictor pre-selection

Predictors were selected based on availability in the NCR. The predictors based on international consensus identified in the Consensus Statement on Mandatory Measurements in Unresectable Pancreatic Cancer Trials (COMM-PACT)[30] served as guidance for the selection. The COMM-PACT predictors include mandatory and recommended baseline and prognostic characteristics that are used in pancreatic cancer studies on systemic treatment to adequately compare outcomes. Depending on availability in the NCR, these factors were used to construct the model, in addition to other potential predictive variables that were available in the NCR. These additional predictive variables were only preselected if they met the following criteria: the number of missing values did not exceed 50% and the variable was not constant across all patients; that is, the variable could potentially improve the model because it discriminated between patients.

Outcome

The outcome of the prediction model is OS, which was measured from the date of diagnosis to the date of death or of last follow-up when the patient was censored.

Model development

The prediction model development followed the same steps as described previously.[31] In short, the following analysis was performed. Cox proportional hazards regression with Efron baseline hazard estimation was used to predict survival. Missing data were multiply imputed (N = 10) according to the MICE (multivariate imputation by chained equations) algorithm during the validation and creation of the final model.[32] Next, predictors were selected from among the preselected variables using the bidirectional Akaike Information Criterion method. Predictors which were selected in a majority of imputations (at least six out of 10), were included in the model. From the resulting model, the concordance index (c-index) and calibration was determined. The c-index is a measure of discriminatory ability and typically ranges from 0.5 (chance level) to 1 (perfect discrimination).[33,34] The calibration refers to the concordance between predicted and observed survival and is displayed in a graph. With perfect calibration, the calibration line has an intercept of 0 and a slope of 1. These methods were used to create the prediction model. To further assess quality, internal–external temporal cross-validation was used, which mimics an external temporal validation.[35] With this validation, a model is created based on earlier diagnosis years and validated based on the subsequent year. In this case, a model was first created based on patients diagnosed

in 2015 and validated based on those diagnosed in 2016. Next, patients from 2015 to 2016 were used to create a model validated based on patients from 2017. Finally, patients diagnosed in the final year, 2018, were used to validate the prediction model based on patients from 2015 to 2017. Analyses were performed using the RStudio environment with R version 3.6.2 (R Foundation for Statistical Computing).

Availability of the data

Data supporting the findings of this study are available from the NCR.[36] Restrictions apply to the availability of these data, which were used under license for this study.

Ethical statements

This report was written in accordance with the TRIPOD guidelines.[37] According to the Central Committee on Research Involving Human Subjects, this type of study does not require approval from an ethics committee in The Netherlands. However, the study was approved by the Privacy Review Board of the NCR and the scientific committee of the Dutch Pancreatic Cancer Group (K18.218).[38]

Results

Patient characteristics

Data on 4,739 patients with metastatic PDAC who were eligible for inclusion were used to create the prediction model. Baseline characteristics of these patients are displayed in Table 1. Of all patients, 48% were women, and the mean age was 69.5 years. Most of the primary tumours were located in the head of the pancreas (41%), followed by the tail (24%) or body (18%). An overlapping lesion was found in 11% of the patients, and in 7% the location was not specified. Three-fourths of the patients had liver metastases (75%), and 26% had peritoneal metastases. Most patients (55%) received BSC only, followed by FOLFIRINOX (19%), biliary drainage only (15%), gemcitabine monotherapy (7%), gemcitabine + nab-paclitaxel (3%), or other systemic treatment (1%).

Characteristic	N (%)
Total, N	4,739
OS, median (IQR), mo	2.5 (2.4–2.6)
Age, mean (SD), y	69.49 (10.26)
BMI, mean (SD), kg/m^2	24.71 (4.10)
Missing	3,322 (70.1)
Sex	
Male	2,444 (51.6)
Female	2,295 (48.4)
WHO performance status	
0	735 (15.5)
1	1,077 (22.7)
2	509 (10.7)
3+	356 (7.5)
Missing	2,062 (43.5)
Albumin, mean (SD), g/L	36.18 (7.84)
Missing	1,517 (32.0)
Bilirubin, mean (SD), µmol/L	65.81 (107.72)
Missing	409 (8.6)
CRP, mean (SD), mg/L	52.18 (65.37)
Missing	906 (19.1)
CA-19.9, mean (SD), kU/L	3,729.31 (3,833.56)
Missing	1,806 (38.1)
Creatinine, mean (SD), µmol/L	75.93 (32.87)
Missing	596 (12.6)
Hemoglobin, mean (SD), mmol/L	7.90 (1.11)
Missing	152 (3.2)
LDH, mean (SD), U/L	281.97 (181.28)
Missing	608 (12.8)
cT stage	
cT_1	185 (3.9)
cT_2	1,206 (25.4)
cT_3	1,314 (27.7)
cT_4	1,406 (29.7)
Missing	628 (13.3)

Table 1 – *continued from previous page*

Characteristic	N (%)
cN stage	
cN_0	1,979 (41.8)
cN_1	1,657 (35.0)
cN_2	223 (4.7)
Missing	880 (18.6)
Tumour topography	
Head of pancreas	1,931 (40.7)
Body of pancreas	835 (17.6)
Tail of pancreas	1,155 (24.4)
Overlapping lesion	499 (10.5)
Pancreas NOS	319 (6.7)
Morphology	
Adenocarcinoma	4,739 (100.0)
Differentiation grade	
G1	41 (0.9)
G2	217 (4.6)
G3	289 (6.1)
Missing	4,192 (88.5)
Distant LN metastasis only	
No	4,590 (96.8)
Yes	151 (3.2)
Liver metastasis	
No	1,201 (25.3)
Yes	3,538 (74.7)
Peritoneal metastasis	
No	3,523 (74.3)
Yes	1,216 (25.7)
Pulmonary metastasis	
No	3,778 (79.8)
Yes	961 (20.2)
Metastatic sites, mean (SD), n	1.55 (0.83)
Type of treatment	
Best supportive care only	2,588 (54.6)
Biliary drainage only	722 (15.2)
FOLFIRINOX	921 (19.4)
Gemcitabine monotherapy	324 (6.8)
Gemcitabine and nab-paclitaxel	120 (2.5)
Other systemic treatment	64 (1.3)

Table 1: Baseline characteristics of included patients.
BMI: body mass index; CRP: C-reactive protein; FOLFIRINOX: 5-FU/leucovorin/oxaliplatin/irinotecan; IQR: interquartile range; LDH: lactate dehydrogenase; LN: lymph node; NOS: not otherwise specified; OS: overall survival.

Model parameters

All possible prognostic variables based on availability in the NCR and variables regarded as mandatory or recommended variables by COMM-PACT are listed in Table 2. Seven COMM-PACT variables were not available in our dataset (i.e., neutrophil-to-lymphocyte ratio, pain at baseline, alkaline phosphatase level, CEA level, previous deep venous thrombosis/embolus, and the global and physical functioning health-related quality of life scales).

Predictors	NCR availability	Included in COMM-PACT	Included in final model
Age	Yes	Yes, mandatory	Yes
Albumin	Yes	Yes, mandatory	Yes
Alkaline phosphatase	Not available	Yes, recommended	No
Bilirubin	Yes	Yes, mandatory	Yes
CA-19.9	Yes	Yes, mandatory	Yes
CEA level	Not available	Yes, recommended	No
cT stage	Yes	Yes, recommended	Yes
cN stage	Yes	Yes, recommended	Yes
CRP	Yes	Yes, mandatory	Yes
Creatinine	Yes	No	No
Distant LN metastasis only	Yes	No	Yes
Hemoglobin	Yes	No	No
LDH	Yes	Yes, mandatory	Yes
Liver metastasis	Yes	Yes, mandatory	Yes
LN metastasis only	Yes	No	No
Neutrophil-to-lymphocyte ratio	Not available	Yes, mandatory	No
Number of metastatic sites	Yes, at most 4 metastatic sites	Yes, mandatory	No
Number of metastatic sites	Yes	No	Yes
Pain at baseline	Not available	Yes, mandatory	No
Performance status	Yes	Yes, mandatory	Yes
Peritoneal metastasis	Yes	Yes, mandatory	Yes
Previous DVT or embolus	Not available	Yes, recommended	No
Primary tumour location	Yes	No	Yes
Pulmonary metastasis	Yes	Yes, recommended	No
HRQoL – global[a]	Not available	Yes, recommended	No
HRQoL – physical functioning[a]	Not available	Yes, recommended	No
Sex	Yes	Yes, recommended	Yes
Type of treatment	Yes	No	Yes

Table 2: Possible prediction model predictors.
COMM-PACT: Consensus Statement on Mandatory Measurements in Unresectable Pancreatic Cancer Trials; CRP: C-reactive protein; DVT: deep venous thrombosis; LDH: lactate dehydrogenase; LN: lymph node; NCR: Netherlands Cancer Registry; HRQoL: Health-Related Quality of life. [a]These factors are relative to the EORTC Quality of Life Questionnaire-Core 30 score.

A total of 16 predictors were selected in the final model and are presented in Table 3 with their accompanying hazard ratios. Model parameters include patient (age, sex, and performance status), laboratory (albumin, C-reactive protein, CA 19-9, lactate dehydrogenase, bilirubin levels), and tumour characteristics (clinical tumour and nodal stage, primary tumour location, distant lymph node metastasis only, liver metastasis, peritoneal metastasis, number of metastatic sites) and treatment strategy. Compared with BSC only, all treatment strategies (i.e., biliary drainage only and systemic treatment strategies) resulted in superior OS. The longest OS was observed in patients who received FOLFIRINOX (hazard ratio, 0.26; 95% CI, 0.24–0.28) (Table 3). An example of predicted and observed risks in 20 patients is displayed in Figure 1.

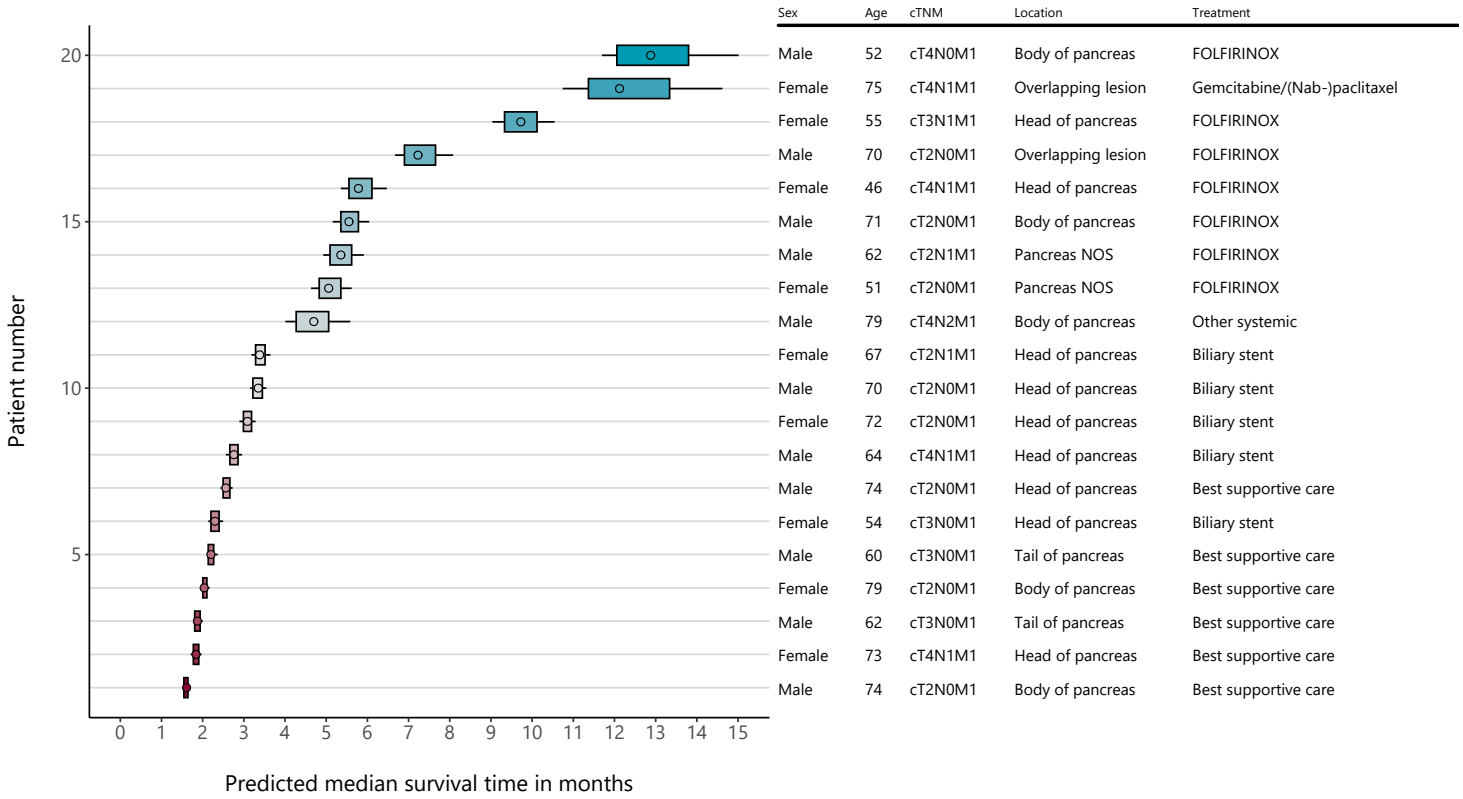


Figure 1: Predicted and observed median survival.
The figure demonstrates the practical applicability of the prediction model in individual patients. The diagram is based on a random sample of 20 patients in the dataset. The bars show the 50% confidence interval and the lines show the 80% confidence interval. The table on the right shows the patient characteristics on which the predictions were based. FOLFIRINOX: 5-FU / leucovorin / oxaliplatin / irinotecan; NOS: not otherwise specified.

Variable	HR (95% CI)
Age, y	1.006 (1.003 – 1.009)
Sex	
Male	1
Female	0.893 (0.843 – 0.947)
WHO performance status	
0	1
1	1.071 (0.994 – 1.155)
2	1.239 (1.137 – 1.350)
3+	1.443 (1.314 – 1.584)
Albumin, g/L	0.983 (0.979 – 0.987)
Bilirubin, µmol/L	1.000 (1.000 – 1.001)
CRP, mg/L	1.003 (1.002 – 1.003)
CA-19.9, kU/L	1.000 (1.000 – 1.000)
LDH, U/L	1.001 (1.000 – 1.001)
Clinical T stage	
cT_1	1
cT_2	1.277 (1.102 – 1.481)
cT_3	1.376 (1.186 – 1.596)
cT_4	1.314 (1.135 – 1.523)
Clinical N stage	
cN_0	1
cN_1	1.118 (1.053 – 1.188)
cN_2	1.277 (1.127 – 1.448)
Primary tumour location	
Head of pancreas	1
Body of pancreas	1.034 (0.947 – 1.129)
Tail of pancreas	1.009 (0.928 – 1.097)
Overlapping lesion	1.044 (0.942 – 1.157)
Pancreas NOS	1.194 (1.055 – 1.351)
Distant LN metastasis only	
No	1
Yes	0.695 (0.580 – 0.833)
Liver metastasis	
No	1
Yes	1.398 (1.290 – 1.515)
Peritoneal metastasis	
No	1
Yes	1.219 (1.126 – 1.321)
Number of metastatic sites	1.069 (1.027 – 1.112)
Type of treatment	
Best supportive care only	1
Biliary drainage only	0.252 (0.230 – 0.276)
FOLFIRINOX	0.694 (0.627 – 0.769)
Gemcitabine monotherapy	0.407 (0.361 – 0.458)
Gemcitabine and nab-paclitaxel	0.269 (0.223 – 0.325)
Other systemic treatment	0.292 (0.226 – 0.378)

Table 3: Overview of model predictors with associated multivariate HRs.
CRP: C-reactive protein; FOLFIRINOX: 5-FU / leucovorin / oxaliplatin / irinotecan; HR: hazard ratio; LDH: lactate dehydrogenase; LN: lymph node; NOS: not otherwise specified.

Performance

Model performance statistics are shown in Table 4. Overall, the model had a good discriminatory ability, with a c-index of 0.72. The model calibration is displayed in Figure 2 and shows an overall good accordance between predicted and observed survival. The calibration is determined at the median OS of 2.5 months after diagnosis. Both the calibration intercept and the slope include the ideal values of 0 and 1 in the 95% confidence interval, respectively.

	Complete model (95% CI)	**Internal-external validation (95% CI)**
C-index	0.73 (0.72 – 0.74)	0.72 (0.71 – 0.73)
Calibration intercept	0.00 (0.00 – 0.00)	0.03 (0.00 – 0.05)
Calibration slope	1.00 (1.00 – 1.00)	0.97 (0.92 – 1.01)
Calibration deviance	0.00 (0.00 – 0.00)	0.01 (0.00 – 0.02)

Table 4: Model performance.

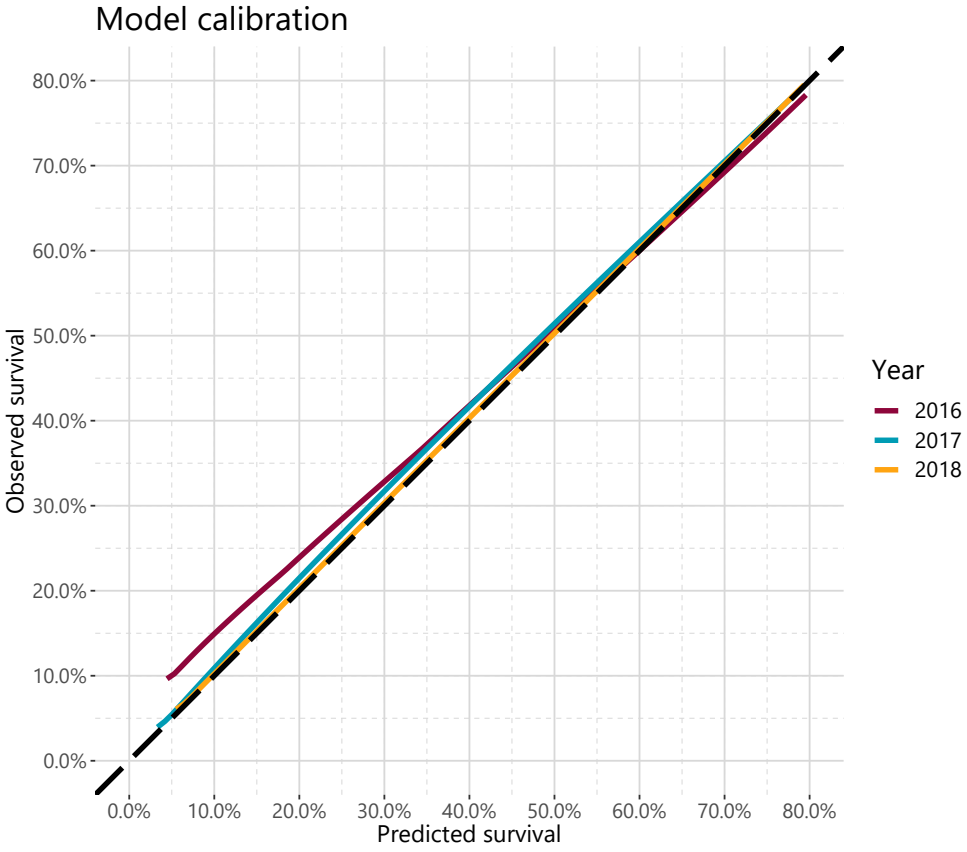


Figure 2: Model calibration.
The figure shows the accordance between survival predicted by the model and observed survival for each validation cohort.

Discussion

This prediction model for patients with synchronous metastatic PDAC is the first model based on a population-based cohort including a nationwide cohort of patients with metastatic disease diagnosed in 2015 through 2018 (N = 4,739) and various types of (systemic) treatments. The SOURCE-PANC model stands out with an applicability to a wide range of patients and good internal–external validation. The model showed good accordance between predicted and observed OS and can be valuable in supporting communication regarding expectations of systemic treatment compared with BSC. The prediction model will be incorporated in a web interface that can be used during consultations in the clinic to contribute to the shared decision-making process. This web interface will be made freely available to the oncologic community and will display personalised survival charts comparing various treatments after input of the model parameters. Results of the exploration of the clinical applicability of an online model for oesophagogastric cancer will be used for the implementation of the model for PDAC (ClinicalTrials.gov identifier: NCT04232735).[31] Prognostic COMM-PACT variables that were identified by experts in the field of pancreatic cancer were added to the model.[30] These possible predictors were selected for use as prognostic variables in randomised controlled trials investigating first-line systemic treatment in unresectable or metastatic pancreatic cancer. The variables include common baseline characteristics that are collected routinely in general clinical practice and are therefore easy to add to a model that could be helpful in predicting OS for various treatments while taking into account these prognostic features. However, not all mandatory and recommended predictors that were defined by the COMM-PACT study were available in the dataset, and some predictors were missing in a considerable number of patients (e.g., performance status, which was missing in 44%). This could have impaired the model's performance.

Major strengths of this study are the inclusion of population-based data that represent all patients with metastatic PDAC in clinical practice and the addition of different systemic treatment regimens to the model. We categorised systemic treatment into the most frequently applied regimens (i.e., FOLFIRINOX[23], gemcitabine + nab-paclitaxel[24], gemcitabine monotherapy[25], or other treatment regimens) The actual survival benefit of the regimens can be assessed in comparison with each other or with BSC only. As a result, patients should consider the possible benefits of FOLFIRINOX in terms of OS against the increased risk of toxicity of this regimen compared with gemcitabine + nab-paclitaxel or gemcitabine monotherapy.[23,39]

With a validated c-index of 0.72, the model can be regarded as adequately discriminative, and the model's c-index is larger than that of most other similar models.[15–22] The model calibration indicates a close coherence between predicted and observed survival. Moreover, validation was performed according to a temporal internal–external scheme resembling a true external validation in future cohorts. External validity of the model with similar cohorts is needed for further verification of the model's accuracy and is crucial to application in clinical practice.[40] Therefore, the next step will be to validate the model in a different population, such as by using more recent data from the NCR or data from the Belgian Cancer Registry, as has been performed previously.[41] Use of more recent data from the NCR would be preferable, because differences in registration practices, interpretations, and missing variables in Belgian data may impair the validation.[41]

This study has some limitations. First, we had information available only on the initial treatment and did not know any therapy beyond first-line treatments. Although we expect only a small number of the patients will eventually be eligible to receive these treatments, treatment options beyond first line are expanding.[42,43] Currently, additional follow-up data are collected in the NCR, and an update of the model that includes second-line systemic treatment strategies will be performed when these data are available. Second, the model focuses only on survival, whereas health-related quality of life (HRQoL) is also of prime importance to these patients[44]. Improvement in HRQoL has been reported for those receiving systemic therapy[45,46], biliary stents[47], and radiotherapy[48], whereas fatigue, pain, and treatment-related toxicity are associated with decreased HRQoL[44,49]. We are currently collecting patient-reported outcome measures on HRQoL and will add this information to the model once a sufficient amount of data has been collected.[50] Third, we did not have data on treatment toxicity, which may be an important factor for patients to consider during treatment decision-making. Furthermore, apart from the number of comorbidities, we did not have data that included comorbidity severity, such as the Charlson comorbidity index. We aim to incorporate these data into the model in the future as well.

The prediction model developed in this study is the first to present OS outcomes in patients with synchronous metastatic PDAC based on a nationwide cohort. SOURCE-PANC can be used to predict OS with good accordance and calibration. Based on this model, a decision support tool will be created to support clinicians in communicating with their patients regarding expectations and prognosis.

References

1. Bertuccio, P., Alicandro, G., Malvezzi, M., *et al.* Cancer mortality in Europe in 2015 and an overview of trends since 1990. Annals of Oncology 30, 1356–1369 (2019).
2. Lepage, C., Capocaccia, R. & Hackl, M. Survival in patients with primary liver cancer, gallbladder and extrahepatic biliary tract cancer and pancreatic cancer in Europe 1999-2007: results of EUROCARE-5. Eur J Cancer 51, 2169–2178 (2015).
3. Baxter, N. N., Whitson, B. A. & Tuttle, T. M. Trends in the treatment and outcome of pancreatic cancer in the United States. Annals of Surgical Oncology 14, 1320–1326 (2007).
4. Van der Geest, L. G., van Eijck, C. H., Groot Koerkamp, B., *et al.* Trends in treatment and survival of patients with nonresected, nonmetastatic pancreatic cancer: A population-based study. Cancer Medicine 7, 4943–4951 (2018).
5. Van der Geest, L. G., Haj Mohammad, N., Besselink, M. G., *et al.* Nationwide trends in chemotherapy use and survival of elderly patients with metastatic pancreatic cancer. Cancer Medicine 6, 2840–2849 (2017).
6. Latenstein, A. E., van der Geest, L. G., Bonsing, B. A., *et al.* Nationwide trends in incidence, treatment and survival of pancreatic ductal adenocarcinoma. European Journal of Cancer 125, 83–93 (2020).
7. Vickers, A. J. Prediction models in cancer care. CA: A Cancer Journal for Clinicians (2011).
8. De Castro, S., Biere, S. S., Lagarde, S. M., *et al.* Validation of a nomogram for predicting survival after resection for adenocarcinoma of the pancreas. British Journal of Surgery 96, 417–423 (2009).
9. Smith, B. J. & Mezhir, J. J. An interactive Bayesian model for prediction of lymph node ratio and survival in pancreatic cancer patients. Journal of the American Medical Informatics Association : JAMIA 21, e203–211 (2014).
10. Haga, Y., Wada, Y., Saitoh, T., *et al.* Value of general surgical risk models for predicting postoperative morbidity and mortality in pancreatic resections for pancreatobiliary carcinomas. Journal of Hepato-Biliary-Pancreatic Sciences 21, 599–606 (2014).
11. Lee, K. H., Lee, J. K., Choi, D. W., *et al.* Postoperative Prognosis Prediction of Pancreatic Cancer With Seven MicroRNAs. Pancreas 44, 764–768 (2015).
12. He, C., Mao, Y., Wang, J., *et al.* Nomograms predict long-term survival for patients with periampullary adenocarcinoma after pancreatoduodenectomy. BMC Cancer 18, 327 (2018).
13. Huang, L., Balavarca, Y., Van der Geest, L., *et al.* Development and validation of a prognostic model to predict the prognosis of patients who underwent chemotherapy and resection of pancreatic adenocarcinoma: A large international population-based cohort study. BMC Medicine 17, 66 (2019).
14. Latenstein, A. E., van Roessel, S., van der Geest, L. G., *et al.* Conditional Survival After Resection for Pancreatic Cancer: A Population-Based Study and Prediction Model. Annals of Surgical Oncology 27, 2516–2524 (2020).
15. Forssell, H., Wester, M., Akesson, K. & Johansson, S. A proposed model for prediction of survival based on a follow-up study in unresectable pancreatic cancer. The BMJ Open 3, e004064 (2013).
16. Vienot, A., Beinse, G., Louvet, C., *et al.* Overall Survival Prediction and Usefulness of Second-Line Chemotherapy in Advanced Pancreatic Adenocarcinoma. Journal of the National Cancer Institute 109 (2017).

17. Hamada, T., Nakai, Y., Yasunaga, H., *et al.* Prognostic nomogram for nonresectable pancreatic cancer treated with gemcitabine-based chemotherapy. British journal of cancer 110, 1943–1949 (2014).
18. Deng, Q., Dong, S. & Wang, L. Development and validation of a nomogram for predicting survival in patients with advanced pancreatic ductal adenocarcinoma. Scientific Reports 7 (2017).
19. Goldstein, D., Von Hoff, D. D., Chiorean, E. G., *et al.* Nomogram for Estimating Overall Survival in Patients with Metastatic Pancreatic Cancer. Pancreas 49, 744–750 (2020).
20. Hang, J., Wu, L., Zhu, L., *et al.* Prediction of overall survival for metastatic pancreatic cancer: Development and validation of a prognostic nomogram with data from open clinical trial and real-world study. Cancer Medicine 7, 2974–2984 (2018).
21. Vernerey, D., Huguet, F., Vienot, A., *et al.* Prognostic nomogram and score to predict overall survival in locally advanced untreated pancreatic cancer (PROLAP). British Journal of Cancer 115, 281–289 (2016).
22. Song, W., Miao, D. L. & Chen, L. Nomogram for predicting survival in patients with pancreatic cancer. OncoTargets and Therapy 11, 539–545 (2018).
23. Conroy, T., Desseigne, F., Ychou, M., *et al.* FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. New England Journal of Medicine 364, 1817–1825 (2011).
24. Von Hoff, D. D., Ervin, T., Arena, F. P., *et al.* Increased Survival in Pancreatic Cancer with nab-Paclitaxel plus Gemcitabine. New England Journal of Medicine 369, 1691–1703 (2013).
25. Burris, H. A., Moore, M. J., Andersen, J., *et al.* Improvements in survival and clinical benefit with gemcitabine as first- line therapy for patients with advanced pancreas cancer: A randomized trial. Journal of Clinical Oncology 15, 2403–2413 (1997).
26. Stiggelbout, A. M., Van der Weijden, T., De Wit, M. P., *et al.* Shared decision making: Really putting patients at the centre of healthcare. The BMJ (Online) 344, e256 (2012).
27. Van de Water, L. F., van Kleef, J. J., Dijksterhuis, W. P., *et al.* Communicating treatment risks and benefits to cancer patients: a systematic review of communication methods. Quality of Life Research 29, 1747–1766 (2020).
28. Fritz, A., Percy, C., Jack, A., *et al.* International Classification of Diseases for Oncology (ICD-O). 3rd ed. Switzerland: World Health Organization. Geneva (2000).
29. Pallialine. Palliatieve sedatie. `https://www.pallialine.nl/sedatie` (2021).
30. Ter Veer, E., van Rijssen, L. B., Besselink, M. G., *et al.* Consensus statement on mandatory measurements in pancreatic cancer trials (COMM-PACT) for systemic treatment of unresectable disease. The Lancet Oncology 19, e151–e160 (2018).
31. Van den Boorn, H. G., Abu-Hanna, A., Ter Veer, E., *et al.* SOURCE: A registry-based prediction model for overall survival in patients with metastatic oesophageal or gastric cancer. Cancers 11, 187 (2019).
32. Van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. Journal of Statistical Software 45, 1–67 (2011).
33. Steyerberg, E. W., Vickers, A. J., Cook, N. R., *et al.* Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology 21, 128–138 (2010).
34. Gerds, T. A. & Schumacher, M. Consistent estimation of the expected brier score in general survival models with right-censored event times. Biometrical Journal 48, 1029–1040 (2006).
35. Steyerberg, E. W. & Harrell, F. E. Prediction models need appropriate internal, internal-external, and external validation. Journal of Clinical Epidemiology 69, 245–247 (2016).
36. Dutch Upper GI Cancer Group. DUCG 2015. `https://www.ducg.nl`.

37. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. BMC Medicine 13 (2015).
38. Strijker, M., Mackay, T. M., Bonsing, B. A., *et al.* Establishing and Coordinating a Nationwide Multidisciplinary Study Group: Lessons Learned by the Dutch Pancreatic Cancer Group. Annals of Surgery 271, E102–E104 (2020).
39. Tahara, J., Shimizu, K., Otsuka, N., *et al.* Gemcitabine plus nab-paclitaxel vs. FOLFIRINOX for patients with advanced pancreatic cancer. Cancer Chemotherapy and Pharmacology 82, 245–250 (2018).
40. Collins, G. S., De Groot, J. A., Dutton, S., *et al.* External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. BMC Medical Research Methodology 14, 40 (2014).
41. Van Kleef, J. J., van den Boorn, H. G., Verhoeven, R. H., *et al.* External validation of the dutch SOURCE survival prediction model in belgian metastatic oesophageal and gastric cancer patients. Cancers 12, 834 (2020).
42. Wang-Gillam, A., Li, C. P., Bodoky, G., *et al.* Nanoliposomal irinotecan with fluorouracil and folinic acid in metastatic pancreatic cancer after previous gemcitabine-based therapy (NAPOLI-1): A global, randomised, open-label, phase 3 trial. The Lancet 387, 545–557 (2016).
43. Oettle, H., Riess, H., Stieler, J. M., *et al.* Second-line oxaliplatin, folinic acid, and fluorouracil versus folinic acid and fluorouracil alone for gemcitabine-refractory pancreatic cancer: Outcomes from the CONKO-003 Trial. Journal of Clinical Oncology 32, 2423–2429 (2014).
44. Lewis, A. R., Pihlak, R. & McNamara, M. G. The importance of quality-of-life management in patients with advanced pancreatic ductal adenocarcinoma. Current Problems in Cancer 42, 26–39 (2018).
45. Glimelius, B., Hoffman, K., Sjödén, P. O., *et al.* Chemotherapy improves survival and quality of life in advanced pancreatic and biliary cancer. Annals of Oncology 7, 593–600 (1996).
46. Gourgou-Bourgade, S., Bascoul-Mollevi, C., Desseigne, F., *et al.* Impact of FOLFIRINOX compared with gemcitabine on quality of life in patients with metastatic pancreatic cancer: Results from the PRODIGE 4/ACCORD 11 randomized trial. Journal of Clinical Oncology 31, 23–29 (2013).
47. Walter, D., Van Boeckel, P. G., Groenen, M. J., *et al.* Higher quality of life after metal stent placement compared with plastic stent placement for malignant extrahepatic bile duct obstruction: A randomized controlled trial. European Journal of Gastroenterology and Hepatology 29, 231–237 (2017).
48. Tian, Q., Zhang, F. & Wang, Y. Clinical assessment of palliative radiotherapy for pancreatic cancer. Cancer/Radiotherapie 22, 778–783 (2018).
49. Carrato, A., Falcone, A., Ducreux, M., *et al.* A Systematic Review of the Burden of Pancreatic Cancer in Europe: Real-World Impact on Survival, Quality of Life and Costs. Journal of Gastrointestinal Cancer 46, 201–211 (2015).
50. Van den Braak, R. R. J. C., van Rijssen, L. B. & van Kleef, J. J. Nationwide comprehensive gastro-intestinal cancer cohorts: the 3P initiative. Acta Oncol 57, 195–202 (2018).

Chapter 7

Health-related quality of life during metastatic systemic therapy for oesophagogastric cancer

A systematic review and meta-analysis

Jessy Joy van Kleef, Emil ter Veer, Héctor G. van den Boorn, Sandor Schokker, Lok Lam Ngai, Mariska J. Prins, Nadia Haj Mohammad, Lonneke V. van de Poll-Franse, Aeilko H. Zwinderman, Martijn G.H. van Oijen, Mirjam A.G. Sprangers and Hanneke W.M. van Laarhoven

Based on:

J.J. van Kleef, E. ter Veer, H.G. van den Boorn, *et al.* Quality of life during palliative systemic therapy for esophagogastric cancer: systematic review and meta-analysis. Journal of the National Cancer Institute 112 (2020).

Abstract

Background

Palliative systemic therapy can prolong life and reduce tumour-related symptoms for patients with advanced oesophagogastric cancer. However, side effects of treatment could negatively affect health-related quality of life (HRQoL). Our aim was to review the literature and conduct a meta-analysis to examine the effect of palliative systemic therapy on HRQoL.

Methods

EMBASE, Medline, and Central were searched for phase II/III randomised controlled trials until April 2018 investigating palliative systemic therapy and HRQoL. A meta-analysis was performed on baseline and follow-up summary values of global health status (GHS) and other European Organisation for Research and Treatment of Cancer scales. A clinically relevant change and difference of 10 points (scale 0–100) was set to assess the course of HRQoL over time within treatment arms as well as between arms.

Results

We included 43 randomised controlled trials (N = 13,727 patients). In the first-line and beyond first-line treatment setting, pooled baseline GHS mean estimates were 54.6 (95% CI: 51.9–57.3) and 57.9 (95% CI: 55.7–60.1), respectively. Thirty-nine (81.3%) treatment arms showed a stable GHS over the course of time. Anthracycline-based triplets, fluoropyrimidine-based doublets without cisplatin, and the addition of trastuzumab to chemotherapy were found to have favourable HRQoL outcomes. HRQoL benefit was observed for taxane monotherapy and several targeted agents over best supportive care beyond first line.

Discussion

Patients reported impaired GHS at baseline and generally remained stable over time. Anthracycline-based triplets and fluoropyrimidine-based doublets without cisplatin may be preferable first-line treatment options regarding HRQoL for HER2-negative disease. Taxanes and targeted agents could provide HRQoL benefit beyond first line compared with best supportive care.

Introduction

Patients with advanced oesophagogastric cancer face a poor prognosis, with a median overall survival (OS) of 3 to 5 months with best supportive care (BSC).[1–3] Palliative chemotherapy offers life-prolongation and may reduce tumour-related symptoms.[4,5] However, for individual patients, benefit of chemotherapy is uncertain and side effects may occur. Preferably, survival gain through chemotherapy should not be attained at the expense of health-related quality of life (HRQoL). The impact of chemotherapy on HRQoL has been investigated by a review of 19 clinical trials up to 2008.[6] More recently published reviews addressed HRQoL as a secondary outcome and summarised findings of randomised controlled trials (RCTs) narratively.[4,5]

In this study, we aimed to examine the impact of systemic therapy on HRQoL of patients with advanced oesophagogastric cancer more comprehensively using meta-analysis to answer the following four research questions: What are the most affected disease-related functions and symptoms before start of treatment in the first-line and beyond first-line treatment setting? What is the course of HRQoL over time? Which chemotherapy regimens show better HRQoL over comparator regimens? Is there a relationship between HRQoL and OS?

Methods

This manuscript was written in accordance with the PRISMA guidelines.[7]

Search strategy

Medline, EMBASE, and the Cochrane Central Register of Controlled Trials were searched for RCTs up to April 2018. Medical subject headings and words for oesophagogastric cancer and several treatment options were used. Details regarding the search strategy can be found in the Supplementary Table 1. Titles, abstracts, and full texts were screened by at least two reviewers. Disagreements were discussed until consensus was reached.

Study selection and quality assessment

Phase II and III RCTs were included if they compared palliative systemic therapies for patients with metastatic, unresectable, or recurrent adenocarcinoma of the stomach or oesophagus and provided information regarding planned HRQoL analyses. Studies using solely self-constructed or non-validated measures were excluded. Study quality was assessed by two reviewers using the Cochrane risk of bias tool (version 5.1.0). Items were scored as unknown, low, or high risk of bias. The 'minimum standard checklist for evaluating HRQoL outcomes in cancer clinical trials'[8] was used to assess the quality of HRQoL reporting and was rated by two reviewers. The procedures of this quality assessment were published earlier.[9] When studies did not elaborate on the clinical significance of their HRQoL findings, this criterion was nonetheless rated as 'satisfied' when a time-to-deterioration analysis (TtD analysis) was performed with a specific threshold. That threshold was interpreted as a clinically meaningful change.

Data extraction

Data were extracted and checked by one reviewer using a pre-formatted Excel document, including the following items: patient characteristics, HRQoL measure used, HRQoL scores at baseline and during follow-up, HRQoL scores between study arms, and OS statistics. We contacted corresponding authors by e-mail for clarification in case of conflicting or unclear reporting. When no response was given, data from figures were chosen over text, and more recently published data were chosen over previously published data in case multiple articles on the same RCT were available. Plot Digitizer 2.6.8 was used to digitise figures to obtain values.

If studies did not report the sample size per measurement, it was linearly interpolated. When studies did not report a hazard ratio regarding OS analysis, it was estimated on the basis of the published survival curve as published previously.[10]

HRQoL measures

Most RCTs investigating HRQoL in oesophagogastric cancer used the cancer-specific QLQ-C30 questionnaire developed and validated by the European Organisation for Research and Treatment of Cancer (EORTC).[9] Its 30 items are combined to form five functioning scales (physical, emotional, role, cognitive, and social functioning), three symptom scales (fatigue, pain, and nausea/vomiting), six single items (dyspnoea, insomnia, appetite loss, constipation, diarrhoea, and financial difficulties), and a global health status (GHS) scale. The GHS scale consists of two items questioning the patients' overall health and quality of life. All scale and single-item scores range from 0 to 100 after linear transformation. Higher scores on the GHS and functioning scales represent higher levels of HRQoL or functioning. Higher scores on symptom scales or items indicate higher levels of symptomatology. The QLQ-C30 is supplemented with disease-specific EORTC modules, for example, the OES18 for oesophageal cancer patients and STO22 for gastric cancer patients. These questionnaires are mainly focused on symptoms related to the specific tumour site (e.g., odynophagia), side effects of treatment (e.g., nausea/vomiting), and additional domains affected by the disease or treatment (e.g., body image).

Statistical analysis

HRQoL was the primary outcome, expressed as the GHS scale for studies using the EORTC QLQ-C30. Secondary outcomes were functioning and symptom scales, for example, fatigue and physical functioning. If GHS scores were similar and one or more other HRQoL symptom or functioning scores were in favour of one arm, then that arm was rated as 'superior' in overall HRQoL. When studies used other measures such as the Functional Assessment of Cancer Therapy (FACT) or EQ-5D, the outcome in the particular RCT representing global HRQoL was chosen to represent overall HRQoL.

To assess the most affected disease-related symptoms and functioning scales at baseline, pooled mean estimates of EORTC scales ranging from 0 to 100 were compared. The five most affected symptoms or functioning scales were selected based on the highest amount of deviation from the optimal score. The course of HRQoL over time within treatment arms was investigated by com-

paring the reported mean scores at baseline and during follow-up. Positive and negative changes greater than or equal to 10 points (or 10% of the total scale for other measures) relative to baseline were interpreted as improved or deteriorated HRQoL, respectively. Changes less than 10 points were interpreted as stable to be consistent with most RCTs that considered a threshold of a 10-point change as clinically meaningful.[11]

Reported mean and median scores at baseline and/or follow-up were meta-analysed in R version 3.5.1 with the *metafor* 2.0–0 package.[12] GHS scores were transformed with the logit function to ensure that the lower and upper bounds (i.e., 0–100) would be preserved during analysis. A linear mixed-effects model was fitted on the data with time, regimen (BSC, singlets, doublets, and triplet therapies), and time-regimen interaction as predictors. Furthermore, the time variable was interpolated with a cubic spline to capture nonlinear effects. The individual studies were added as random effects. Treatment regimen effect and change in HRQoL were tested with an omnibus Wald test, followed by a post hoc test in case of statistically significant results.

To assess HRQoL between treatment arms, a 10-point difference was used irrespective of statistical significance. If no statistical comparison was reported but study reports concluded that HRQoL differences between arms were present or absent, we interpreted the results as such. When other analyses were performed, for example, TtD analysis or responder analysis, a given treatment arm was deemed favourable if results were statistically different ($p < 0.05$). Therapies were compared on the basis of the following drug classes as reported previously[13]: BSC, singlets (fluoropyrimidine monotherapy), doublets (cisplatin-based doublets and fluoropyrimidine-non-cisplatin–based doublets), and triplets (anthracycline-based triplets and taxane-based triplets).

A tetrachoric correlation coefficient (two-sided) was calculated to assess the degree of association between differential HRQoL and differential OS within an RCT. Two dichotomous classifications were made based on RCT comparisons in which the experimental arm was rated as 'improved' or 'similar' with regard to the OS hazard ratio ($p < 0.05$). HRQoL of the experimental arm was also rated as 'improved' or 'similar' in comparison with the control arm. Studies were weighted according to the total number of patients in the treatment arms.

Results

Search results

In total, 43 unique RCTs were included (N = 13,727); see Supplementary Figure 1. Nine studies published HRQoL findings separately.[14–22] Thirty-one studies investigated HRQoL in the first-line treatment setting (N = 9,214)[3,23–52], and 12 studies investigated HRQoL beyond the first-line treatment setting (N = 4,513)[53–64]. Baseline characteristics of the included studies are shown in Table 1.

Study quality

Study quality according to the Cochrane risk of bias tool is shown in Supplementary Table 2. Twenty-eight (65.1%) studies were rated as low risk of bias, and 15 (34.9%) studies were rated as unclear on at least one item. The quality of HRQoL reporting is presented in Supplementary Table 3. Eleven studies were rated as 'probably robust,' 27 as 'limited,' and five as 'very limited.'

HRQoL measures

Most studies (N = 39, 90.7%) used the EORTC QLQ-C30 to assess HRQoL (Table 1). EORTC disease-specific modules were always used in addition to the QLQ-C30. The FACT questionnaires (FACT-Biologic Response Modifier, FACT-Gastric module) were used in two studies. The EQ-5D was used in eight studies and often in combination with the QLQ-C30.

Clinically meaningful differences

In total, 24 studies (55.8%) did not describe a threshold for clinically meaningful differences (or change). Eleven studies chose a 10-point difference, seven studies a 5-point difference, and one study an 8-point difference. Three studies performed sensitivity analyses with other threshold values (often with 5 and 10 points, but also with 15, 20, and 30) yielding similar results.[14,17,22]

Study	N	Arms	N Men (%)	Median age, y (range)	N metastatic (%)	N WHO PS $\geq$ 2 (%)	N OES (%)	N GEJ (%)	N GAS (%)	HRQoL measure
First-line treatment										
Ajani et al., 2010[23]; Bodoky et al., 2015[18]	521	Cis + S-1	382 (73.3)	59 (18–83)	497 (95.4)	0 (0)	0 (0)	82 (15.7)	438 (84.1)	FACT-Ga
	508	Cis + 5-FU	347 (68.3)	60 (20–85)	488 (96.1)	0 (0)	0 (0)	88 (17.3)	417 (82.1)	
Al-Batran et al., 2013[24] – Kripp et al., 2014[17]	72	DTX + Ox +5-FU/Lv	51 (70.8)	69 (NA–81)	50 (69.4)	5 (6.9)	0 (0)	27 (37.5)	45 (62.5)	C30, STO22
	71	Ox + 5-FU/Lv	45 (63.4)	70 (NA–82)	49 (69.0)	6 (8.5)	0 (0)	24 (33.8)	47 (66.2)	
Bang et al., 2010[25]; Satoh et al., 2014[22]	294	Trastuzumab + Cis + Cap/5-FU	226 (76.9)	59.4 (10.8)*	284 (96.6)	30 (10.2)	0 (0)	58 (19.7)	236 (80.3)	C30, STO22
	290	Cis + Cap/5-FU	218 (75.2)	58.5 (11.2)*	280 (96.6)	27 (9.3)	0 (0)	48 (16.6)	242 (83.4)	
Bouché et al., 2004[26]	45	5-FU/Lv	37 (82.2)	64 (45–75)	45 (100)	12 (26.7)	0 (0)	13 (28.9)	32 (71.1)	C30
	44	Cis + 5-FU/Lv	35 (79.5)	64 (43–76)	44 (100)	11 (25.0)	0 (0)	13 (29.5)	31 (70.5)	
	45	Iri + 5-FU/Lv	38 (84.4)	65 (37–76)	45 (100)	10 (22.2)	0 (0)	15 (33.3)	30 (66.7)	
Bramhall et al., 2002[27]	185	Marimastat	131 (70.8)	68	136 (73.5)	0 (0)	0 (0)	0 (0)	185 (100)	C30
	184	PLB	131 (71.2)	68	132 (71.7)	0 (0)	0 (0)	0 (0)	184 (100)	
Cunningham et al., 2008[28]	249	Epi + Cis + 5-FU	202 (81.1)	65 (22–83)	198 (79.5)	29 (11.6)	87 (34.9)	72 (28.9)	90 (36.1)	C30
	241	Epi + Cis + Cap	194 (80.5)	64 (25–82)	185 (76.8)	30 (12.4)	71 (29.5)	68 (28.2)	102 (42.3)	
	235	Epi + Ox + 5-FU	191 (81)	61 (33–78)	181 (77.0)	20 (8.5)	93 (39.6)	55 (23.4)	87 (37.0)	
	239	Epi + Ox + Cap	198 (82.8)	62 (25–80)	181 (75.7)	24 (10.0)	82 (34.3)	53 (22.2)	104 (43.5)	
Dank et al., 2008[29]; Curran et al., 2009[15]	170	Iri + 5-FU/Lv	125 (73.5)	58 (29–76)	160 (94.1)	1 (0.6)[a]	0 (0)	34 (20.0)	136 (80.0)	C30, EQ-5D
	163	Cis + 5-FU	108 (66.3)	59 (28–77)	155 (95.1)	2 (1.2)[a]	0 (0)	31 (19.0)	132 (81.0)	
Glimelius et al., 1997[3]	31	Etoposide + 5-FU/Lv	23 (74.2)	64 (45–75)	–	79 (50–100)[bc]	0 (0)	9 (29.0)	22 (71.0)	C30
	30	BSC	22 (73)	63 (40–74)	–	87 (50–100)[bc]	0 (0)	7 (23.3)	23 (76.7)	
Gubanski et al., 2010[30] – Gubanski et al., 2014[16]	39	DTX + 5-FU/Lv	26 (66.7)	63 (39–79)	34 (87.2)	1 (2.6)	0 (0)	39 (100)		C30, STO22
	39	Iri + 5-FU/Lv	34 (87.2)	64 (42–75)	34 (87.2)	5 (12.8)	0 (0)	39 (100)		
Guimbaud et al., 2014[31] – Nuemi et al., 2015[19]	209	Epi + Cis + Cap	154 (73.7)	61 (28–84)	173 (82.8)	36 (17.2)	0 (0)	73 (34.9)	133 (63.9)	C30, STO22
	207	Iri + 5-FU/Lv	155 (74.9)	61 (29–81)	176 (85.0)	27 (13.0)	0 (0)	63 (30.4)	138 (66.7)	
Hall et al., 2017[32]	17	Epi + Ox + Cap	13 (76.5)	74 (64–82)	17 (100)	6 (35.3)	5 (29.4)	2 (11.8)	10 (58.8)	C30, OG25, EQ-5D
	19	Ox + Cap	13 (68.4)	77 (50–85)	17 (89.5)	5 (26.3)	11 (57.9)	1 (5.3)	5 (26.3)	
	19	Cap	15 (78.9)	75 (57–87)	18 (94.7)	7 (36.8)	8 (42.1)	4 (21.1)	7 (36.8)	
Hecht et al., 2016[33]	249	Lapatinib + Ox + Cap	189 (75.9)	61 (19–86)	236 (94.8)	21 (8.4)	12 (4.8)	23 (9.2)	214 (85.9)	C30, OES18, STO22
	238	Ox + Cap	176 (73.9)	59 (27–84)	227 (95.4)	22 (9.2)	8 (3.4)	20 (8.4)	210 (88.2)	
Hwang et al., 2017[34]	26	Cap	16 (61.5)	77 (70–83)	–	6 (23.1)	0 (0)	0 (0)	26 (100)	C30
	24	Ox + Cap	18 (75.0)	75 (70–84)	–	4 (16.7)	0 (0)	0 (0)	24 (100)	
Kim et al., 2012[35]	65	Ox + S-1	44 (67.7)	60 (28–77)	47 (72.3)	0 (0)	0 (0)	0 (0)	65 (100)	C30
	64	Ox + Cap	45 (70.3)	61 (20–75)	46 (71.9)	2 (3.1)	0 (0)	0 (0)	64 (100)	
Kim et al., 2018[36]	53	S-1	39 (73.6)	72 (65–81)	53 (100)	9 (17.0)	0 (0)	0 (0)	53 (100)	C30, STO22
	54	Cap	44 (81.5)	71 (65–78)	54 (100)	9 (16.7)	0 (0)	0 (0)	54 (100)	
Lu et al., 2018[37]	160	PTX + Cap	115 (71.9)	57 (11.3)*	146 (91.3)	86.4 (7.0) *[c]	0 (0)	68 (42.5)	92 (57.5)	C30, STO22
	160	Cis + Cap	118 (73.8)	56 (10.9)*	140 (87.5)	86.1 (7.4) *[c]	0 (0)	63 (39.4)	97 (60.6)	

Table 1 – *continued from previous page*

Study	N	Arms	N Men (%)	Median age, y (range)	N metastatic (%)	N WHO PS $\geq$ 2 (%)	N OES (%)	N GEJ (%)	N GAS (%)	HRQoL measure
Ohtsu et al., 2011[38]	387	Bevacizumab + Cis + Cap	257 (66.4)	58 (22–81)	367 (94.8)	22 (5.7)	0 (0)	54 (14.0)	333 (86.0)	C30, ST022
	387	Cis + Cap	258 (66.7)	59 (22–82)	378 (97.7)	20 (5.2)	0 (0)	49 (12.7)	338 (87.3)	
Park et al., 2006[39]	38	PTX + 5-FU	19 (50.0)	53 (36–73)	38 (100)	8 (21.1)	0 (0)	0 (0)	38 (100)	C30
	39	DXT + 5-FU	26 (66.7)	51 (27–74)	39 (100)	6 (15.4)	0 (0)	0 (0)	39 (100)	
Park et al., 2007[40]	45	Iri + 5-FU/Lv	30 (66.7)	55 (26–73)	–	11 (24.4)	0 (0)	0 (0)	45 (100)	C30
	45	PTX + Iri + 5-FU/Lv	30 (66.7)	51 (29–70)	–	7 (15.6)	0 (0)	0 (0)	45 (100)	
Park et al., 2017[41]	59	S-1 + Ox continuous	37 (62.7)	54 (28–68)	59 (100)	1 (1.7)	0 (0)	0 (0)	59 (100)	C30, STO22
	62	S-1 + Ox stop-and-go	44 (71)	53 (29–69)	62 (100)	1 (1.6)	0 (0)	0 (0)	62 (100)	
Rao et al., 2010[42]	35	Matuzumab + Epi + Cis + Cap	24 (68.6)	59 (29–79)	35 (100)	0 (0)	0 (0)	21 (60.0)	14 (40.0)	C30, OES18
	36	Epi + Cis + Cap	27 (75.0)	64 (36–76)	36 (100)	0 (0)	0 (0)	20 (55.6)	16 (44.4)	
Ross et al., 2002[43]	289	Epi + Cis + 5-FU	218 (75.4)	58 (28–78)	154 (53.3)	60 (20.8)	95 (32.9)	61 (21.1)	128 (44.3)	C30
	285	MMC + Cis + 5-FU	226 (79.3)	59 (29–77)	174 (61.1)	49 (17.2)	93 (32.6)	64 (22.5)	120 (42.1)	
Roth et al., 2007[44]	40	Epi + Cis + 5-FU	30 (75.0)	59 (32–71)	33 (82.5)	0 (0)	0 (0)	0 (0)	40 (100)	C30
	38	DTX + Cis	29 (76.3)	58 (40–70)	31 (81.6)	0 (0)	0 (0)	0 (0)	38 (100)	
	41	DTX + Cis +5-FU	30 (73.2)	61 (35–78)	39 (95.1)	0 (0)	0 (0)	0 (0)	41 (100)	
Ryu et al., 2015[45]	306	S-1 + Cis (3w)	231 (75.5)	60 (27–74)	303 (99.0)	10 (3.3)	0 (0)	306 (100)		EQ-5D
	309	S-1 + Cis (5w)	233 (75.4)	59 (29–74)	306 (99.0)	3 (1.0)	0 (0)	309 (100)		
Sadighi et al., 2006[46]	42	Epi + Cis + 5-FU	34 (81.0)	57.2 (9.8)*	42 (100)	–	0 (0)	0 (0)	42 (100)	C30
	44	DTX + Cis + 5-FU	31 (70.5)	55.4 (14.0)*	44 (100)	–	0 (0)	0 (0)	44 (100)	
Tebbutt et al., 2002[47]	123	5-FU	94 (76.4)	72 (52–84)	71 (57.7)	37 (30.1)	29 (23.6)	33 (26.8)	55 (44.7)	C30
	127	5-FU + MMC	95 (74.8)	72 (52–84)	73 (57.5)	44 (34.6)	27 (21.3)	30 (23.6)	69 (54.3)	
Tebbutt et al., 2010[48]	50	DTX + Cis + 5-FU	42 (84.0)	60.5 (11.5)*	49 (98.0)	1 (2.0)	11 (22.0)	13 (26.0)	26 (52.0)	C30, OES18, STO22
	56	DTX + Cap	42 (75.0)	59.1 (10.8)*	56 (100)	2 (3.6)	20 (35.7)	13 (23.2)	23 (41.1)	
Tebbutt et al., 2016[49]	37	Panitumumab + DTX + Cis + Cap/5-FU	33 (89.2)	64 (40–79)	37 (100)	2 (5.4)	15 (40.5)	10 (27.0)	13 (35.1)	C30, OES18, STO22
	39	DTX + Cis + Cap/5-FU	30 (76.9)	59 (37–77)	39 (100)	3 (7.7)	13 (33.3)	11 (28.2)	15 (38.5)	
Van Cutsem et al., 2006[50]; Ajani et al., 2007[14]	221	DTX + Cis + 5-FU	159 (71.9)	55 (26–79)	213 (96.4)	3 (1.4)[a]	0 (0)	42 (19.0)	179 (81.0)	C30, EQ-5D
	224	Cis + 5-FU	158 (70.5)	55 (25–76)	217 (96.9)	3 (1.3)[a]	0 (0)	56 (25.0)	168 (75.0)	
Webb et al., 1997[51]	126	Epi + Cis + 5-FU	99 (78.6)	59 (35–79)	79 (62.7)	30 (23.8)	27 (21.4)	27 (21.4)	72 (57.1)	C30
	130	Doxo + 5-FU + MTX	110 (84.6)	60 (29–78)	79 (60.8)	32 (24.6)	24 (18.5)	33 (25.4)	73 (56.2)	
Yoshino et al., 2016[52]	146	S-1	107 (73.3)	74 (32–94)	–	5 (3.4)	0 (0)	0 (0)	146 (100)	FACT-BRM
	149	S-1 + lentinan	101 (67.8)	73 (44–93)	–	8 (5.4)	0 (0)	0 (0)	149 (100)	
Beyond first-line treatment										
Bang et al., 2017[53]	263	Olaparib + PTX	174 (66.2)	58 (49–67)	260 (99.0)	0 (0)	0 (0)	12 (5.0)	250 (95.0)	C30, STO22
	262	PLB + PTX	185 (70.6)	59 (50–65)	254 (96.9)	0 (0)	0 (0)	2 (0.8)	259 (98.9)	
Dutton et al., 2014[54]	224	Gefitinib	183 (81.7)	65 (58–70)	–	50 (22.3)	171 (76.3)	53 (23.7)	0 (0)	C30, OG25
	225	PLB	189 (84.0)	65 (58–71)	–	44 (19.6)	181 (80.4)	44 (19.6)	0 (0)	
Ford et al., 2014[55]	84	DTX + BSC	69 (82.1)	65 (29–84)	73 (86.9)	14 (16.7)	18 (21.4)	27 (32.1)	39 (46.4)	C30, STO22
	84	BSC	67 (79.8)	66 (36–84)	74 (88.1)	12 (14.3)	15 (17.9)	32 (38.1)	37 (44.0)	

Table 1 – *continued from previous page*

Study	N	Arms	N Men (%)	Median age, y (range)	N metastatic (%)	N WHO PS $\geq$ 2 (%)	N OES (%)	N GEJ (%)	N GAS (%)	HRQoL measure
Fuchs et al., 2014[56]	238	Ramucirumab	169 (71.0)	60 (52–67)	–	0 (0)	0 (0)	60 (25.2)	178 (74.8)	C30
	117	BSC	79 (67.5)	60 (51–71)	–	1 (0.9)	0 (0)	30 (25.6)	87 (74.4)	
Lee et al., 2017[64]	23	DTX	18 (78.3)	56 (34–68)	22 (95.7)	0 (0)	0 (0)	0 (0)	23 (100)	C30, STO22
	23	DTX + Cis	20 (87.0)	55 (38–74)	18 (78.3)	2 (8.7)	0 (0)	0 (0)	23 (100)	
	23	DTX + S-1	14 (60.9)	55 (39–68)	18 (78.3)	2 (8.7)	0 (0)	0 (0)	23 (100)	
Li et al., 2013[57]	47	Apatinib 850 mg once daily	39 (83.0)	55 (NA)	43 (91.5)	0 (0)	0 (0)	47(100)		C30
	46	Apatinib 425 mg twice daily	34 (73.9)	53 (NA)	45 (97.8)	0 (0)	0 (0)	46 (100)		
	48	PLB	36 (75.0)	54 (NA)	48 (100)	0 (0)	0 (0)	48 (100)		
Li et al., 2016[58]	176	Apatinib 850 mg once daily	132 (75.0)	58 (32–71)	–	0 (0)	0 (0)	22 (21.8)	69 (68.3)	C30
	91	PLB + BSC	69 (75.8)	58 (28–70)	–	0 (0)	0 (0)	14 (23.3)	43 (71.7)	
Ohtsu et al., 2013[59]	439	Everolimus	322 (73.3)	62 (20–86)	439 (100)	25 (5.7)	0 (0)	118 (26.9)	321 (73.1)	C30
	217	PLB + BSC	161 (74.2)	62 (20–88)	217 (100)	27 (12.4)	0 (0)	69 (31.8)	148 (68.2)	
Pavlakis et al., 2016[60]; Martin et al., 2018[21]	97	Regorafenib + BSC	78 (80.4)	63 (33–81)	96 (99.0)	0 (0)	0 (0)	37 (38.1)	60 (61.9)	C30, STO22, EQ-5D
	50	PTX + BSC	40 (80.0)	62 (32–85)	48 (96.0)	0 (0)	0 (0)	19 (38.0)	31 (62.0)	
Shitara et al., 2017[61]	243	N-PTX 3w	178 (73.3)	66 (60–72)	–	4 (1.6)	0 (0)	0 (0)	243 (100)	EQ-5D
	240	N-PTX weekly	178 (74.2)	67 (60–72)	–	2 (0.8)	0 (0)	0 (0)	240 (100)	
	243	S-PTX weekly	176 (72.4)	65 (59–71)	–	4 (1.6)	0 (0)	0 (0)	243 (100)	
Thuss-Patience et al., 2017[62]	228	Trastuzumab emtansine	177 (77.6)	62 (19–79)	218 (95.6)	1 (0.4)	0 (0)	77 (33.8)	151 (66.2)	C30, STO22, EQ-5D
	117	PTX/DTX	95 (81.2)	62 (27–80)	113 (96.6)	1 (0.9)	0 (0)	33 (28.2)	84 (71.8)	
Wilke et al., 2014[63]; Al-Batran et al., 2016[20]	330	Ramucirumab + PTX	229 (69.4)	61 (25–83)	–	0 (0)	0 (0)	66 (20.0)	264 (80.0)	C30, EQ-5D
	335	PTX + PLB	243 (72.5)	61 (24–84)	–	0 (0)	0 (0)	71 (21.2)	264 (78.8)	

Table 1: Baseline characteristics of included studies.

Data derived from original study article based on entire analysis-based sample. *Mean and SD. OES: oesophagus; GEJ: gastro-oesophageal junction; GAS: stomach; BSC: best supportive care; C30: EORTC QLQ-C30; Cap: capecitabine; Cis: cisplatin; Doxo: doxorubicin; DTX: docetaxel; EORTC: European Organisation for Research and Treatment of Cancer; Epi: epirubicin; FACT: Functional Assessment of Cancer Therapy; FACT-BRM: FACT-Biologic Response Modifier, FACT-Ga FACT-Gastric cancer subscale; 5-FU: 5 fluorouracil; Iri: irinotecan; Lv: leucovorin; MMC: mitomycin; NA: not available; N-PTX: nab-paclitaxel; OES18: EORTC QLQ-OES18; OG25: EORTC QLQ-OG25; Ox: oxaliplatin; PLB: placebo; PTX: paclitaxel; S-1: tegafur/gimeracil plus oteracil; S-PTX: solvent-based paclitaxel; STO22: EORTC QLQ-STO22; WHO PS: World Health Organisation Performance Status; 3w: three times per week; 5w: five times per week. [a]Karnofsky Performance Status 70. [b]Mean and range. [c]Karnofsky Performance Status.

Baseline HRQoL

Thirteen of 31 studies (41.9%) investigating first-line therapies reported HRQoL scores at baseline. Mean GHS at baseline ranged from 43.0 to 67.9. Meta-analysis showed a pooled mean GHS of 54.6 (95% CI = 51.9 to 57.3); see Figure 1. Mean baseline estimates of other HRQoL scales are shown in Supplementary Figure 2. The five most affected EORTC scales at baseline were anxiety, GHS, fatigue, appetite loss, and pain in the stomach area. Five of 12 studies (41.7%) investigating beyond first-line therapies reported HRQoL scores at baseline. Mean GHS at baseline ranged between 43.6 and 61.5. Meta-analysis showed a pooled mean GHS of 57.9 (95% CI = 55.7 to 60.1); see Figure 1. Supplementary Figure 3 shows the mean baseline estimates of other HRQoL scales. Anxiety, GHS, fatigue, appetite loss, and weight loss were the most affected HRQoL scales.

Mean HRQoL scores over time

In total, 22 RCTs with a total of 48 study arms reported on the course of HRQoL over time. Thirty-nine (81.3%) treatment arms showed a stable GHS over the course of time.

Sixteen RCTs investigating first-line treatments with a total of 34 study arms reported on longitudinal HRQoL. Twenty-eight arms showed stable, five arms showed improved, and one arm showed deteriorated HRQoL over a short period ($<$18 weeks); see Supplementary Figure 4. Mixed-model analysis of follow-up GHS data showed no statistically significant time effect ($\beta = 0.00$ per day, p = 1.00). In addition, no differences in GHS were found between first-line treatments groups; BSC, singlets, doublets and triplets (Wald z-statistic = 3.31, p = 0.35). However, the analysis showed some improvement in GHS of 11.9 points for singlets in the long term ($\pm$4 months); see Figure 2. Other HRQoL scales showed a clinically significant improvement greater than or equal to 10 points over time in pain (-16.8 points), stomach pain (-16.2 points), appetite loss (-15.4 points), eating restrictions (-10.9 points), dysphagia (-10.6 points), and emotional functioning (11.4 points). However, patients scored higher on 'upset by hair loss' (13.4 points); see Supplementary Figures 5.

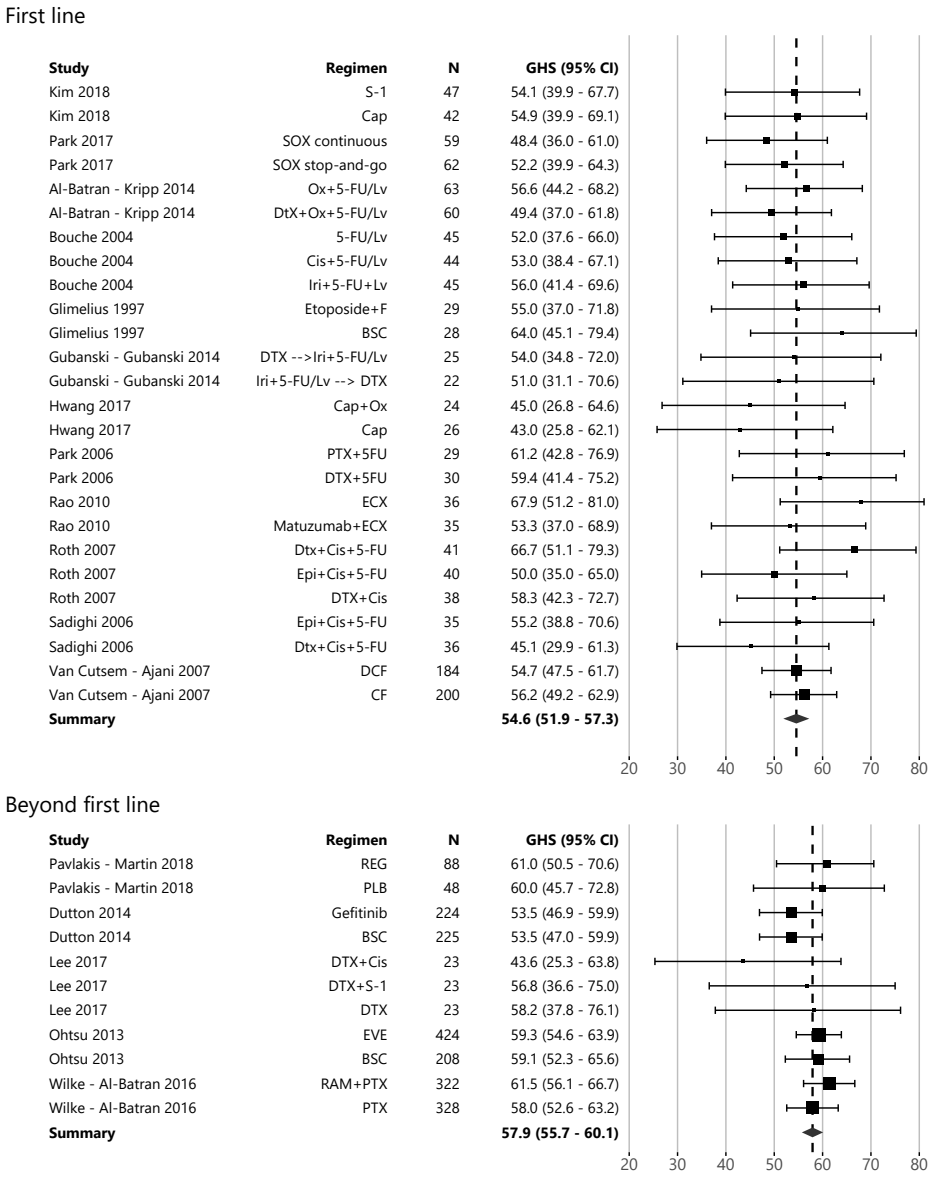


Figure 1: Forest plots of patient-reported baseline global health status (GHS).
Squares indicate mean GHS summary scores with a 95% confidence interval. The size of the squares indicate the weight of the study in the meta-analysis. The summary statistic is the result of a random-effects meta-analysis with a logit transformation applied to the study score. Diamonds indicate the 95% confidence interval of the summary statistic. BSC: best supportive care; Cap: capecitabine; CF: cisplatin and 5-FU; Cis: cisplatin; DCF: docetaxel, cisplatin, and 5-FU; DTX: docetaxel; ECX: epirubicin, cisplatin, and capecitabine; Epi: epirubicin; EVE: everolimus; 5-F: 5-fluorouracil; Iri: irinotecan; Lv: leucovorin; Ox: oxaliplatin; PLB: placebo; PTX: paclitaxel; RAM: ramucirumab; REG: regorafenib; S-1: tegafur/gimeracil plus oteracil; SOX: S-1 and oxaliplatin.

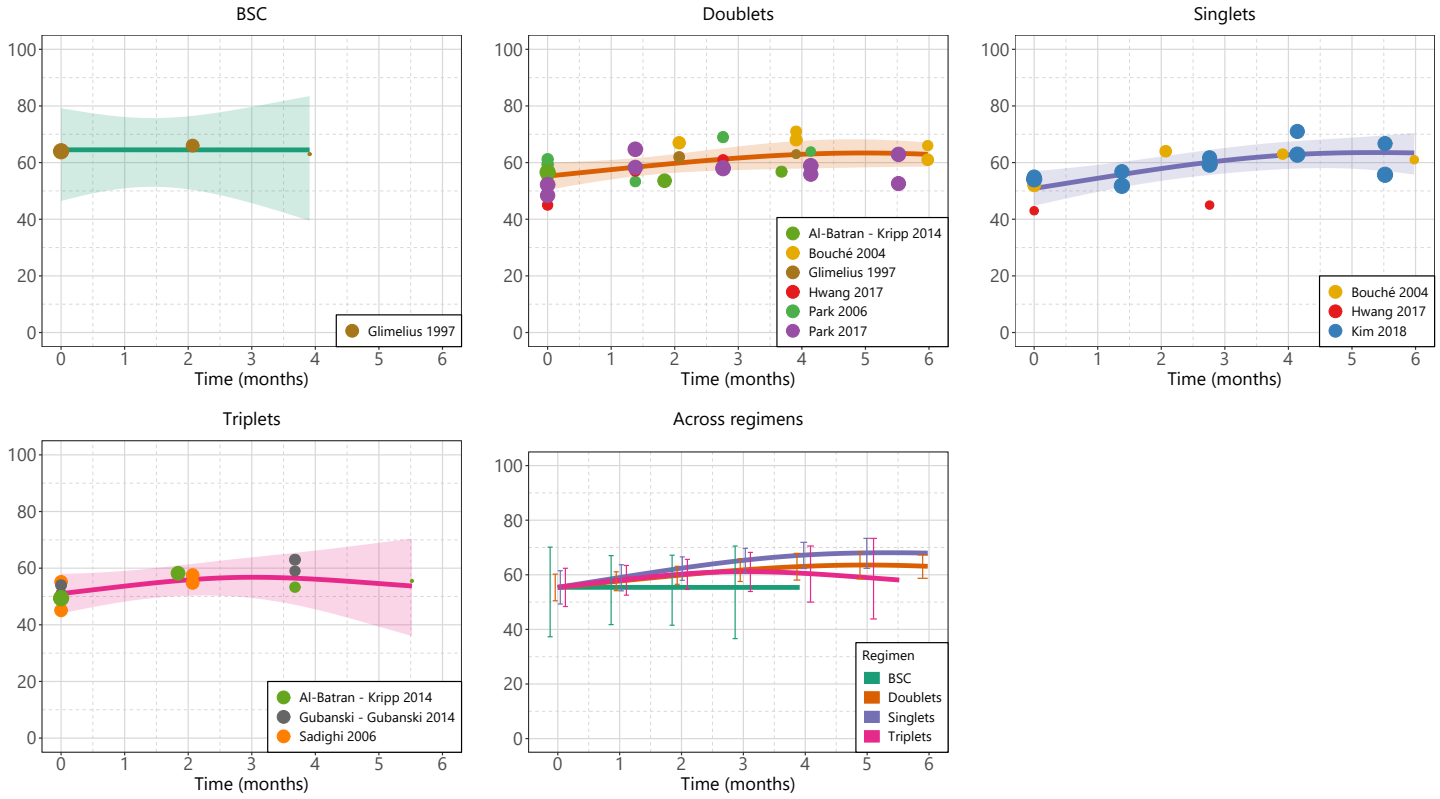


Figure 2: Global health status (GHS) during first-line therapy.
BSC: GHS estimates with regard to best supportive care. Singlets: GHS estimates with regard to singlet therapy. Doublets: GHS estimates with regard to doublet therapy. Triplets: GHS estimates with regard to triplet therapy. Across regimes: combination of the other panels. GHS estimates were derived from linear mixed effects modelling. Error bars and coloured bands indicate 95% confidence interval. Coloured dots indicate specific study arms. Larger dots indicate larger sample sizes. BSC: best supportive care.

Six RCTs investigating beyond first-line treatments with a total of 14 study arms reported on the course of HRQoL over time. Eleven arms showed stable, one arm showed improved, and two arms showed deteriorated HRQoL over a short period (<18 weeks). Mixed-model analysis showed no time effect (β = -0.00 per day, p = 0.31) of GHS and stayed within a 10-point difference relative to baseline; see Figure 3. No statistically significant treatment-time interaction of BSC vs. singlets and doublets (Wald z-statistic = 0.22, p = 0.64) was observed beyond first line. Other HRQoL scales showed a clinically significant worsening greater than 10 points in role functioning (-12.0 points), fatigue (11.7 points), appetite loss (16.4 points), and being upset by hair loss (13.7 points); see Supplementary Figures 6.

HRQoL differences between treatments

Of the 37 comparisons made between first-line treatment regimens, most studies (N = 30) reported similar GHS (Table 2). Only six comparisons showed a superior GHS favouring one particular arm. Of those six, four arms consisted of the anthracycline-based triplet epirubicine, cisplatin, and 5-FU (ECF). Notably, when other HRQoL scales were taken into account, almost one-half of the first-line studies (20 of 37) showed a superior HRQoL.

Two first-line studies compared capecitabine and oxaliplatin (CAPOX) with capecitabine (Cap), and both showed superior overall HRQoL in CAPOX-treated patients.[32,34] Other doublets compared with singlets did not show this clinically significant result, except for irinotecan and 5-FU/leucovorin (Lv) vs. 5-FU/Lv.[26]

Fluoropyrimidine-based doublets (without cisplatin) showed comparable results to cisplatin-based doublets regarding GHS but showed favourable outcomes on several other HRQoL scales in two first-line phase III RCTs.[29,37] One other phase II trial showed differences between mean scores on almost all HRQoL scales in favour of fluoropyrimidine-based doublets, but these differences were below the 10-point threshold.[26]

When comparing first-line anthracycline-based triplets with fluoropyrimidine-based doublets (without cisplatin), one phase III and one phase II trial reported similar outcomes in terms of HRQoL and OS.[31,32] A cisplatin-based doublet showed worse GHS scores compared with an anthracycline-based triplet.[44]

First-line targeted therapies vs. chemotherapy only showed similar overall GHS between arms, but differences were found on other HRQoL scales (Table 2). TtD analysis in the 'Trastuzumab for Gastric Cancer' study showed a prolonged time to deterioration in all scales of the QLQ-C30 and QLQ-STO22 for patients treated with trastuzumab, cisplatin, and capecitabine or FU.[22]

The effect of a targeted agent on HRQoL vs. BSC was investigated in six RCTs beyond the first-line treatment setting. GHS scores were comparable between targeted agents and BSC. Patients treated with ramucirumab reported more often (34%) improved or stable GHS than patients treated with BSC (13%). This difference was not statistically significant.[56] When other HRQoL scales were taken into account, most targeted agents showed better overall HRQoL compared with BSC (Table 3). Two studies investigated the effect of a targeted agent in addition to taxane-monotherapy beyond first line.[20,53] Time to GHS deterioration greater than or equal to 10 points was similar between arms. However, in the RAINBOW trial, ramucirumab plus paclitaxel affected emotional functioning and nausea or vomiting favourably but diarrhoea adversely in the TtD analysis. Responder analysis also showed favourable outcomes for the ramucirumab plus paclitaxel arm with regard to GHS, physical and role functioning, pain, fatigue, and appetite loss. EQ-5D scores were found to be comparable.[20]

The effect of chemotherapy without targeted agents on HRQoL vs. BSC was assessed in one phase III RCT investigating docetaxel beyond first line.[55] GHS scores were comparable between arms, but the following scales or items favoured docetaxel: dysphagia, pain, abdominal pain, constipation, and nausea or vomiting.

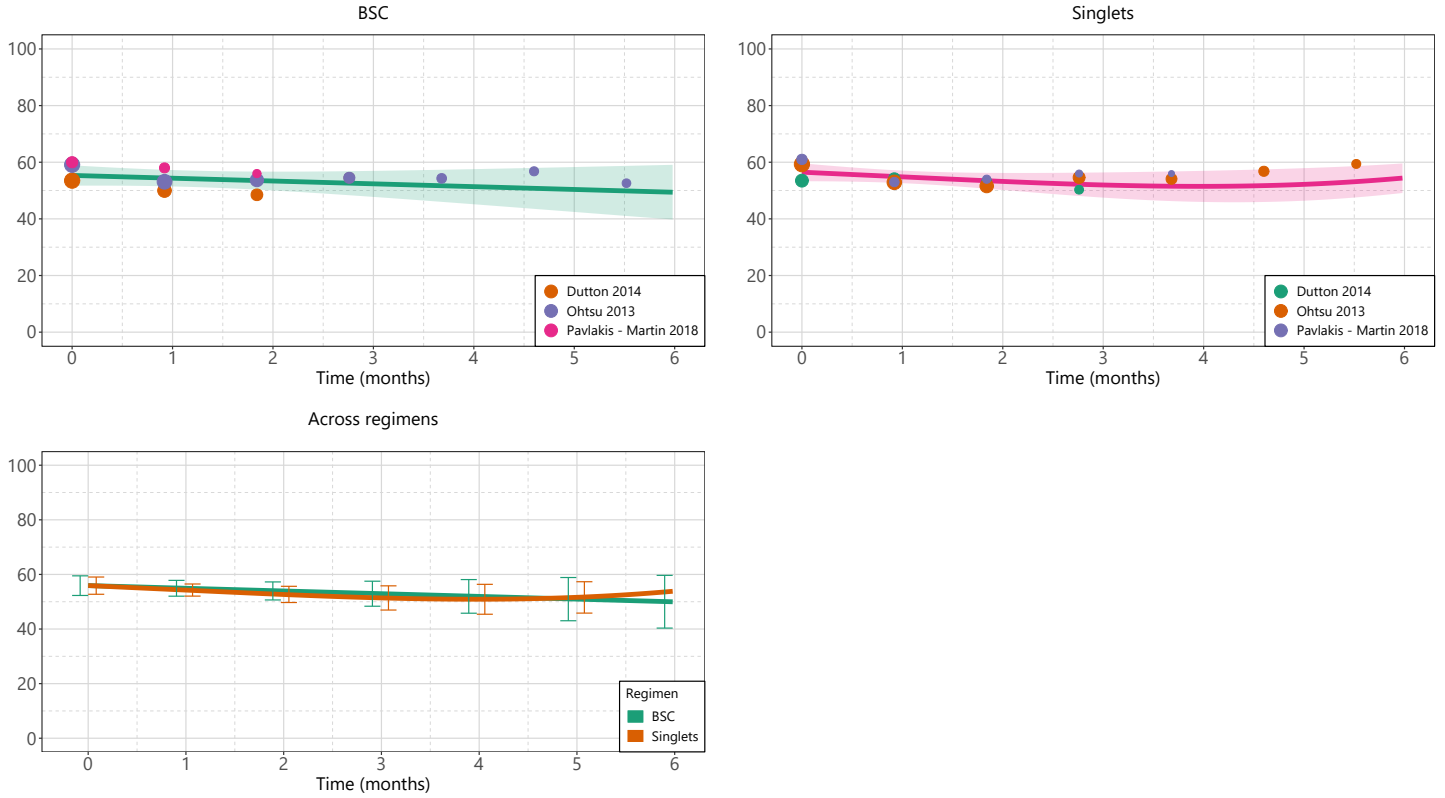


Figure 3. Global health status (GHS) during beyond first-line therapy.
BSC: GHS estimates with regard to best supportive care, or placebo. Singlet: estimates with regard to singlet therapy. Across regimes: combination of the other panels. GHS estimates were derived from linear mixed-effects modelling. Error bars and coloured bands indicate 95% confidence interval. Coloured dots indicate specific study arms. Larger dots indicate larger sample sizes. BSC: best supportive care.

Time-to-deteriation analysis

TtD analysis was conducted in 12 RCTs. The definition of 'deterioration' varied across studies as well as the threshold that defined a clinically meaningful change. Most studies defined an event as a 'definite' deterioration, that is, a decrease of at least 10 points in HRQoL with no subsequent improvement relative to baseline[14,17,21,22,29,31,59], whereas few other studies defined an event as the first decrease greater than or equal to 10 points in HRQoL.[18,20,35,37,53] Additionally, one study included disease progression and treatment discontinuation.[21]

Median time to GHS deterioration of at least 10 points ranged from 2.4 to 4.9 months during first-line systemic therapy without targeted agents. An overview of median times to deterioration are given in Tables 2 and 3 for patients treated in the first line and the beyond first-line setting, respectively.

HRQoL and treatment efficacy

A combined analysis of regimens in both treatment lines showed a weak positive correlation between differential overall HRQoL and differential OS (tetrachoric correlation coefficient: ρ =0.274, N = 49, p = 0.006); see Tables 2 and 3. Our analysis showed no statistically significant correlation when superior HRQoL was defined on the basis of GHS only and no other HRQoL scales were taken into account (ρ =-0.158, N = 49, p = 0.10).

Study	Regimen	TtD GHS 10 points	TtD GHS 5 points	GHS	Differences in other HRQoL scales and items	Overall HRQoL	OS
Doublet-F vs. BSC							
Glimelius et al., 1997[3]	BSC	–	–	Referent	Superior appetite loss (-19), physical (+17), role (+14), and social (+13) functioning at 4 months	Referent	
	5-FU/Lv + etoposide	–	–	Similar (+9)	Superior cognitive functioning (+10, +12) and nausea/vomiting (-27, -12) at 2 and 4 months, respectively, and pain (-10) at 4 months	Superior	Superior
Singlet-F vs. another Singlet-F							
Kim et al., 2018[36a]	Cap	–	–	Referent	Superior nausea/vomiting (-16, -13[b]) at 6 and 12 weeks, physical functioning (+15[c], +11[c]) at 12 and 18 weeks, dyspnoea (-11[c]) at 18 weeks, appetite loss (-11[c]) at 12 weeks, financial difficulties (-11[c]) at 18 weeks, dry mouth (-25[c]) at 12 weeks, and body image (+25[c]) at 12 weeks	Referent	
	S-1	–	–	Similar (-7 [c])	Superior emotional functioning (+12[b]) at 12 and 18 weeks, social functioning (+13[c]) at 12 weeks, pain (-12[c], -16[c]) at 12 and 18 weeks, insomnia (-12[c]) at 12 weeks, constipation (-11[c],-14[c]) at 6 and 12 weeks, and taste (-13[c]) at 18 weeks	Similar	Similar
Singlet-F vs. Doublet-F							
Hwang et al., 2017[34]	Cap	–	–	Referent	Superior role functioning (+19, +18) at 6 and 12 weeks	Referent	
	Cap + Ox	–	–	Superior (+14)	Superior fatigue (-11) and constipation (-10) at 12 weeks and insomnia (-12) at 6 weeks	Superior	Similar
Hall et al., 2017[32]	Cap	–	–	Referent	–	Referent	
	Ox + Cap	–	–	Similar (+6)	Superior fatigue (-17) at 12 weeks	Superior	Superior
Tebbutt et al., 2002[47]	5-FU	–	–	Referent	–	Referent	
	Mitomycin + 5-FU	–	–	Similar	–	Similar	Similar
Yoshino et al., 2016[52]	S-1	–	–	Referent	–	Referent	
	Lentinan + S-1	–	–	Similar	–	Similar	Similar
Bouché et al., 2004[26]	5-FU/Lv	–	–	Referent	–	Referent	
	Iri + 5-FU/Lv	–	–	Similar (+2[c])	Superior constipation (-12[b]), fatigue (-10[c]), and insomnia (-10[c])	Superior	Superior
Singlet-F vs. Doublet-C							
Bouché et al., 2004[26]	5-FU/Lv	–	–	Referent	–	Referent	
	Cis + 5-FU/Lv	–	–	Similar	–	Similar	Superior
Doublet-C vs. another Doublet-C							
Ajani et al., 2010[23]; Bodoky et al., 2015[18]	5-FU + Cis	–	–	Referent	Worse median time to worsening of physical well-being of 3 months	Referent	
	S-1 + Cis	–	–	Similar	Superior median time to worsening of physical well-being of 4.5 months[b]	Superior	Similar
Ryu et al., 2015[45]	S-1/Cis 5w	–	–	Referent	Superior mobility (5% and 11% lesser deterioration at week 6 and 12)[b]	Referent	
	S-1/Cis 3w	–	–	–	Superior pain/discomfort (7% and 8% lesser deterioration at week 12 and 24)[b]	Similar	Similar

Table 2 – *continued from previous page*

Study	Regimen	TtD GHS 10 points	TtD GHS 5 points	GHS	Differences in other HRQoL scales and items	Overall HRQoL	OS
Doublet-F vs. another Doublet-F							
Park et al., 2017[41a]	S-1 + OX continuous	–	–	Referent	–	Referent	
	S-1 + OX stop-and-go	–	–	Similar	Superior role functioning (+11[b]) at week 30 and 36	Superior	Similar
Park et al., 2006[39]	DTX + 5-FU	–	–	Referent	Superior cognitive (+17) and emotional (+11) functioning, and nausea/vomiting (-10)	Referent	
	PTX + 5-FU	–	–	Similar	Superior pain (-10) and appetite loss (-20)	Worse	Similar
Kim et al., 2012[35]	Ox + Cap	4.9 (3.2–6.6)	–	Referent	–	Referent	Referent
	Ox + S-1	4.3 (1.2–7.4)[c]	–	Similar	–	Similar	Similar
Doublet-F vs. Doublet-C							
Lu et al., 2018[37]	Cis + Cap	–	–	Referent	–	Referent	
	PTX + Cap	–	–	Similar	Superior TtD NA% on nausea/vomiting[b], appetite loss[b], financial difficulties[b], reflux,[b] and eating restrictions[b]	Superior	Similar
Dank et al., 2008[29] Curran et al., 2009[15]	Cis + 5-FU	–	5.9 (4.8–7.7)	Referent		Referent	
	Iri + 5-FU/Lv	–	4.9 (3.7–7)[c]	Similar	Superior EQ-5D utility index[b]	Superior	Similar
Bouché et al., 2004[26]	Cis + 5-FU/Lv	–	–	Referent	–	Referent	
	Iri + 5-FU+Lv	–	–	Similar (+1[c])	–	Similar	Similar
Doublet-F vs. Triplet-T							
Al-Batran et al. 2013[24]; Kripp et al., 2014[17]	Ox + 5-FU/Lv	3.9 (2.1–5.6)	3.9 (2.1–6.1)	Referent	–	Referent	
	DTX + Ox+5-FU/Lv	2.4 (2.1–4.8)[c]	2.4 (2.1–4.8)[c]	Similar	–	Similar	Similar
Doublet-C vs. Triplet-T							
Van Cutsem et al., 2006[50]; Ajani et al., 2007[14]	Cis + 5-FU	–	4.2	Referent	–	Referent	
	DTX + Cis+5-FU	–	6.5 HR 0.69 (0.52–0.93)[b]	Superior [b]	Superior TtD 10% analysis on physical[b] and social functioning[b], nausea/vomiting[b], pain,[b] and EQ-5D VAS[b]	Superior	Superior
Roth et al., 2007[44]	DTX + Cis	–	–	Referent	Superior emotional functioning (+13) at 6 weeks	Referent	
	DTX + Cis + 5-FU	–	–	Similar (+8)	Worse role functioning (-17[b]) at 18 weeks, superior constipation (-17) at 18 weeks	Worse	–

Table 2 – *continued from previous page*

Study	Regimen	TtD GHS 10 points	TtD GHS 5 points	GHS	Differences in other HRQoL scales and items	Overall HRQoL	OS
Triplet-A vs. another Triplet-A							
Cunningham et al., 2008[28]	Epi + Cis + 5-FU	–	–	Referent	–	Referent	
	Epi + Cis + Cap	–	–	Similar	–	Similar	Similar
	Epi + Ox + 5-FU	–	–	Similar	–	Similar	Similar
	Epi + Ox + Cap	–	–	Similar	–	Similar	Superior
Webb et al., 1997[51]	5-FU + Doxo + methotrexate	–	–	Referent	–	Referent	
	Epi + Cis + 5-FU	–	–	Superior[d]	–	Superior	Superior
Triplet-A vs. Singlet-F							
Hall et al., 2017[32]	Cap	–	–	Referent	–	Referent	
	Epi + Ox + Cap	–	–	Similar (+3)	Superior fatigue (-11) at 12 weeks	Superior	Similar
Triplet-T vs. another Triplet-T							
Gubanski et al., 2010[30]; Gubanski et al., 2014[16]	Iri + 5-FU/Lv + DTX	–	–	Referent	Superior role (+15[c]) functioning and appetite loss (-10[c])	Referent	
	DTX + Iri + 5-FU/Lv	–	–	Similar	Superior emotional (+10[c]), and social (+13[c]) functioning, diarrhoea (-12[c])	Superior	Similar
Triplet-T vs. Triplet-A							
Sadighi et al., 2006[46]	Epi + Cis + 5-FU	–	–	Referent	–	Referent	
	DTX + Cis + 5-FU	–	–	Similar	–	Similar	Similar
Roth et al., 2007[44]	DTX + Cis + 5-FU	–	–	Referent	Worse role functioning (-17[b]) at 12 and 18 weeks, superior constipation (-17) at 18 weeks	Referent	
	Epi + Cis + 5-FU	–	–	Superior (+17[b])	–	Superior	–
Doublet-C vs. Triplet-A							
Roth et al., 2007[44]	DTX + Cis	–	–	Referent	Superior emotional functioning (+14) at week 6	Referent	
	Epi + Cis + 5-FU	–	–	Superior (+21[b])	–	Superior	–
Triplet-C vs. Triplet-A							
Ross et al., 2002[43]	Mitomycin + Cis + 5-FU	–	–	Referent	–	Referent	
	Epi + Cis + 5-FU	–	–	Superior [d]	Superior physical[d], emotional[d], and cognitive[d] functioning and fatigue[d]	Superior	Similar

Table 2 – *continued from previous page*

Study	Regimen	TtD GHS 10 points	TtD GHS 5 points	GHS	Differences in other HRQoL scales and items	Overall HRQoL	OS
Doublet-F vs. Triplet-A							
Guimbaud et al., 2014[31]; Nuemi et al., 2015[19]	Iri + 5-FU/Lv	–	7.4 (6.2–8.6)	Referent	–	Referent	
	Epi + Cis + Cap	–	7.6 (6.1–8.9)[c]	Similar	–	Similar	Similar
Hall et al., 2017[32]	Ox + Cap	–	–	Referent	–	Referent	
	Epi + Ox + Cap	–	–	Similar (-2.5)	–	Similar	Similar
Doublet-F vs. Triplet-C							
Park et al., 2008[40]	Cis + Iri + 5-FU/Lv	–	–	Referent	–	Referent	
	Iri + 5-FU/Lv	–	–	Similar	–	Similar	Similar
Monoclonal antibodies							
Bang et al., 2010[25]; Satoh et al., 2014[22a]	Cis + Cap/5-FU	6.4	5.6	Referent	–	Referent	
	Trastuzumab + Cis + Cap/5-FU	10.2[b]	8.9[b]	Similar	Superior TtD 10% on all C30 and STO22 scales[b]	Superior	Superior
Ohtsu et al., 2011[38]	Cis + Cap	–	–	Referent	–	Referent	
	Bevacizumab + Cis + Cap	–	–	Similar	–	Similar	Similar
Rao et al., 2010[42]	Epi + Cis + Cap	–	–	Referent	–	Referent	
	Matuzumab + Epi + Cis + Cap	–	–	Similar	–	Similar	Similar
Other targeted therapies							
Bramhall et al., 2002[27]	PLB	–	–	Referent	–	Referent	
	Marimastat	–	–	Similar	–	Similar	Superior
Hecht et al., 2016[33]	Ox + Cap	–	–	Referent	Superior diarrhoea[d] (over first 30 weeks)	Referent	
	Lapatinib + Ox + Cap	–	–	Similar	Superior role[d] and cognitive[d] functioning, nausea/vomiting[d], and constipation[d]	Superior	Similar

Table 2: Treatment comparisons in the first-line treatment setting.

Time to deterioration is given in months and/or stated as a hazard ratio with a corresponding 95% confidence interval. Superior HRQoL refers to a better HRQoL with regard to the comparator arm. For symptom scales it refers to a lesser burden of that symptom. For functioning scales it refers to better functioning. Worse HRQoL refers to a poorer HRQoL with regard to the comparator arm. For symptom scales it refers to a bigger burden of that symptom. For functioning scales it refers to poorer functioning. Similar HRQoL refers to similar HRQoL between the two treatment arms. –: no differences in HRQoL or that the study did not report or perform HRQoL analyses on the topic in question. A: anthracycline; BSC: best supportive care; C30: EORTC QLQ-C30; Cap: capecitabine; Cis: cisplatin; Doxo: doxorubicin; EORTC: European Organisation for Research and Treatment of Cancer; Epi: epirubicin; DTX: docetaxel; F: fluoropyrimidine; 5-FU: 5-fluorouracil; GHS: global health status; HR: hazard ratio; HRQoL: composite measure of GHS and other scales; Iri: irinotecan; Lv: leucovorin; NA: not available; OS: overall survival; Ox: oxaliplatin; PLB: placebo; PTX: paclitaxel; S-1: tegafur/gimeracil plus oteracil; STO22: EORTC QLQ-STO22; T: taxane; TtD: time to deterioration; VAS: visual analog scale; 3w: three times per week; 5w: five times per week. [a]HRQoL was assessed until 18 weeks after randomisation. [b]HRQoL differences between treatment arms were statistically significantly different. [c]HRQoL differences between treatment arms were not statistically significantly different. [d]Statistically significant differences ($p < 0.05$) between treatment arms were reported, but no further data were provided.

Study	Regimen	TtD GHS 10 points	TtD GHS 5 points	GHS	Differences in other HRQoL scales and items	Overall HRQoL	OS
Targeted therapies vs. BSC							
Ohtsu et al., 2013[59]	BSC	–	1.45	Referent	–	Referent	
	Everolimus	–	1.51 HR 0.84 (0.69–1.03)[c]	Similar	–	Similar	Similar
Pavlakis et al., 2016[60] Martin et al., 2018[21]	PLB	Referent	–	Referent	Superior diarrhoea (-13[b]) at 4 weeks	Referent	
	Regorafenib	HR 0.53 (0.37–0.75)[b]?	–	Similar	Superior pain (-10[c]) and abdominal pain (-10[c]) both at week 8. TtD 10% on physical functioning HR 0.50 (0.35–0.72)[b]?	Superior	Similar
Li et al., 2013[57]	PLB	–	–	Referent	–	Referent	
	Apatinib 850 mg	–	–	Similar	Superior insomnia[a] at 8 weeks	Superior	Superior
	Apatinib 425 mg	–	–			Superior	Superior
Li et al., 2016[58]	BSC	–	–	Referent	–	Referent	
	Apatinib 850 mg	–	–	Similar	–	Similar	Superior
Dutton et al., 2014[54]	BSC	–	–	Referent	Superior diarrhoea (+19[b]) at 4 weeks	Referent	
	Gefitinib	–	–	Similar (+3[c])	Superior speech (-10[b]), constipation (-15[b]), and hair loss (-14) at 4 weeks	Superior	Similar
Fuchs et al., 2014[56]	BSC	–	–	Referent	–	Referent	
	Ramucirumab	–	–	Difference of 21% of patients more stable/improved at 6 weeks[c]	–	Superior	Superior
Targeted therapies added to taxane vs. taxane monotherapy							
Wilke et al., 2014[63]; Al-Batran et al., 2016[20]	PLB + PTX	Referent	–	Referent	Superior TtD 10% on diarrhoea[b] HR 1.33 (1.01–1.76)[b]	Referent	
	Ramucirumab + PTX	HR 0.93 (0.73–1.18)[c]	–	A larger proportion of patients experienced stable/improved GHS over time[b]	Superior TtD 10% on emotional functioning HR 0.64 (0.49–0.84)[b], and nausea/vomiting HR 0.75 (0.57–0.9)[b]. Larger proportion of patients experienced stable/improved physical[b] and role functioning[b], pain[b], fatigue[b], and appetite scores[b]	Superior	Superior
Bang et al., 2017[53]	PLB + PTX	2.4	–	Referent	–	Referent	
	Olaparib + PTX	3.4 HR 0.82 (0.64–1.05)[c]	–	Similar	–	Similar	Similar
Taxane-containing chemotherapy vs. BSC							
Ford et al., 2014[55]	BSC	–	–	Referent	–	Referent	
	DTX	–	–	Similar	Superior dysphagia[b], pain[b], abdominal pain[b], constipation[b], and nausea/vomiting[b]	Superior	Superior

Table 3 – *continued from previous page*

Study	Regimen	TtD GHS 10 points	TtD GHS 5 points	GHS	Differences in other HRQoL scales and items	Overall HRQoL	OS
Taxane-containing chemotherapies vs. other therapies							
Thuss-Patience et al., 2017[62]	DTX/PTX	–	–	Referent	–	Referent	
	Trastuzumab emtansine	–	–	Similar	–	Similar	Similar
Shitara et al., 2017[61]	3w N-PTX	–	–	Referent	–	Referent	
	Weekly N-PTX	–	–	Superior	–	Superior	Similar
	Weekly S-PTX	–	–	Superior	–	Superior	Similar
Lee et al., 2017[64]	DTX	–	–	Referent	Not assessed[e]	Referent	
	DTX + Cis	–	–	Similar[c]		Not assessed[e]	Worse
	DTX + S-1	–	–	Similar[c]			Similar

Table 3. Treatment comparisons in beyond first-line treatment setting.

Time to deterioration is given in months and/or stated as a hazard ratio with a corresponding 95% confidence interval. Superior HRQoL refers to a better HRQoL with regard to the comparator arm. For symptom scales it refers to a lesser burden of that symptom. For functioning scales it refers to better functioning. Worse HRQoL refers to a poorer HRQoL with regard to the comparator arm. For symptom scales it refers to a bigger burden of that symptom. For functioning scales it refers to poorer functioning. Similar HRQoL refers to similar HRQoL between the two treatment arms. -—: no differences in HRQoL or that the study did not report or perform HRQoL analyses on the topic in question. BSC: best supportive care; Cis: cisplatin; DTX: docetaxel; GHS: global health status; HR: hazard ratio; HRQoL: composite measure of GHS and other scales; N-PTX: nab-paclitaxel; OS: overall survival ; PLB: placebo; PTX: paclitaxel; S-1: tegafur/gimeracil plus oteracil; S-PTX: solvent-based paclitaxel; TtD: time to deterioration; 3w: three times per week. [a]Statistically significant differences ($p < 0.05$) between treatment arms were reported, but no further data were provided. [b]HRQoL differences between treatment arms were statistically significantly different. [c]HRQoL differences between treatment arms were not statistically significantly different. [d]The time-to-deterioration analysis also included disease progression, treatment discontinuation, and death as an event. [e]Sample size was too small.

Discussion

We reviewed the impact of systemic therapy on HRQoL of patients with advanced oesophagogastric cancer. Our aim was fourfold: first, to gain insight into HRQoL of patients before the start of treatment; second, to investigate the course of HRQoL over time; third, to assess which chemotherapy regimens showed better HRQoL over comparator regimens; and fourth, to assess the relationship between HRQoL and OS.

First, our findings indicate that before the start of chemotherapy, patients reported impaired HRQoL. In the first-line treatment setting, scores were mostly impaired for anxiety, GHS, fatigue, appetite loss, and stomach pain. Beyond first line, patients also reported to be worried about their weight loss. Regarding our primary endpoint, meta-analysis showed that patients reported mean GHS scores of 54.6 and 57.9 for the first-line and beyond first-line setting, respectively. These scores are similar to the EORTC reference values for patients with oesophageal and gastric cancer: 55.6 and 53.1, respectively.[65] These reference values, however, are based on a mixed sample of patients staged I–IV, suggesting that the stage IV patients in the included RCTs experienced relatively good HRQoL. The same holds for patients treated beyond the first line, given the similar GHS scores compared with first line.

Second, GHS scores remained stable during treatment in the majority of RCTs in both lines of treatment. A previous review on this topic supports this finding.[6] When taking into account other HRQoL scales during treatment, improvement was observed for various symptoms (e.g., pain, stomach pain, appetite loss, eating restrictions, and dysphagia) and emotional functioning in the first-line treatment setting. Beyond first line, deterioration in HRQoL was observed for role functioning, fatigue, and being upset by hair loss. Our data show that although GHS might be rated by the patient as stable, changes in other HRQoL domains may be present and therefore not captured or underestimated during trial participation when GHS is the only endpoint of interest. One could also argue that first-line treatments may benefit patients given the stable or improved course of HRQoL over time. However, beyond first line, treatment benefit may be limited given that HRQoL was either stable or deteriorated. Those treatments, however, did show benefit over BSC, suggesting that deterioration is inevitable but may be slowed when treated with systemic therapy in comparison with BSC.

Unfortunately, only one-half of the included study arms (48 of 94) reported on change over time. Consequently, reporting bias may be present where inferior HRQoL in the experimental vs. the control arm may have led to the omission of reporting this finding. Additionally, reporting bias could also be present in the mixed-model analysis over time, in which studies with improved HRQoL might report more details during follow-up. The results of the mixed-model analysis should therefore be interpreted with caution.

Third, GHS comparisons between treatment regimens showed generally no major differences, except for the anthracycline-based triplet ECF. When functioning and symptom scales were taken into account, more differences between arms could be observed with regard to HRQoL. This finding underscores the importance of including the more sensitive and informative functioning and symptom scales in addition to GHS to assess the impact of systemic therapies on HRQoL.[66] Although anthracycline-based triplets showed better HRQoL results than cisplatin-based doublets and cisplatin-based triplets, and fluoropyrimidine-based doublets without cisplatin leaned toward better HRQoL than cisplatin-based doublets, our findings should be interpreted with caution given the limited amount of evidence available.

Fourth, our analysis showed that differential GHS did not correlate with differential OS. A review on colorectal cancer showed similar results where GHS did not differ between arms with a differential primary outcome (including OS, progression-free survival, and response rate).[66] However, we did find a positive weak correlation ($\rho = 0.274$) when other HRQoL subscales were considered. Three RCTs included in this review investigated the relationship between HRQoL and tumour response based on their individual patient data. All three reported that patients with tumour response more often were found to have a stable or improved GHS compared with patients with no tumour response.[16,20,46] One might argue that individual patient data are more appropriate to test this hypothesis given the increased level of detail and precision.

It should be noted that the HRQoL analyses are based on scores obtained from patients who are still alive and able to participate in RCTs and to complete questionnaires. This selective retention may introduce optimism bias where results are not generalisable to the general patient population. In this study, a clinically meaningful difference or change was set at 10 points. Recent literature has highlighted that thresholds are dependent on the scale of interest, and for change they are additionally dependent on the baseline value and the direction of change. The heterogeneity of the included studies did not allow

for differentiating between thresholds. It is reassuring to know that a 10-point difference can be interpreted as clinically relevant according to Cocks and colleagues.[67,68]

As reported previously, the quality of HRQoL outcome reporting is limited in the field of oesophagogastric cancer.[9] A limitation of our analysis is that we did not account for the (lack of) quality, given the limited amount of data and studies to be compared. If only high-quality studies were analysed, the reduction in data would probably hinder the possibility to generate new hypotheses. Standardisation of methods, statistics, and reporting is needed to enable reliable comparisons and (network) meta-analyses.

To accomplish this, we recommend future studies to investigate other HRQoL aspects besides GHS, use supplementary site- and/or treatment-specific questionnaires to detect potential benefits and harms of treatments, and employ the CONSORT PRO statement from trial development to final reporting.[69] Standardisation of statistical analyses with, for example, linear mixed-modelling techniques for repeated measures would also strengthen the reliability of trial results.[70] In addition, where RCTs are mainly focused on head-to-head comparisons of treatments, they are also valuable data sources of HRQoL changes over time.

Change in HRQoL is also valuable information to disclose in the consultation room. Although treatment differences with regard to HRQoL may be modest, patients still should be informed (if they wish) about what to expect when facing the decision to undergo systemic therapy.[71] Our findings suggest that survival gain by means of chemotherapy does not necessarily come at the expense of HRQoL. Therefore, possible preconceptions of patients favouring BSC to maximise their HRQoL should be addressed.

Because the decision for a certain systemic regimen should not be based solely on the effect on HRQoL or efficacy or toxicity, a decision has to be made that takes all outcomes into account. Previous evidence showed that anthracycline-based triplets do not prolong OS and progression-free survival over fluoropyrimidine-based doublets without cisplatin. However, the latter regimen has a more convenient toxicity profile.[13] Therefore, one might consider fluoropyrimidine-based doublets without cisplatin as a preferable first-line treatment option when taking HRQoL, efficacy, and toxicity all into account.

In conclusion, patients reported impaired HRQoL at baseline, which generally remained stable during systemic therapy. Based on the current evidence, anthracycline-based triplets and fluoropyrimidine-based doublets without cisplatin may be preferable first-line treatment options regarding HRQoL. Taxanes and targeted agents could benefit HRQoL beyond first line compared with BSC. Our findings could enable shared decision making during doctor-patient consultations, where the impact of systemic therapy on survival, side effects, and HRQoL are discussed.

References

1. Murad, A. M., Santiago, F. F., Petroianu, A., *et al.* Modified therapy with 5-fluorouracil, doxorubicin, and methotrexate in advanced gastric cancer. Cancer 72, 37–41 (1993).
2. Pyrhönen, S., Kuitunen, T., Nyandoto, P. & Kouri, M. Randomised comparison of fluorouracil, epidoxorubicin and methotrexate (FEMTX) plus supportive care with supportive care alone in patients with non-resectable gastric cancer. British Journal of Cancer 71, 587–591 (1995).
3. Glimelius, B., Ekström, K., Hoffman, K., *et al.* Randomized comparison between chemotherapy plus best supportive care with best supportive care in advanced gastric cancer. Annals of Oncology 8, 163–168 (1997).
4. Wagner, A. D., Syn, N. L., Moehler, M., *et al.* Chemotherapy for advanced gastric cancer. Cochrane Database of Systematic Reviews 2017 (2017).
5. Janmaat, V. T., Steyerberg, E. W., van der Gaast, A., *et al.* Palliative chemotherapy and targeted therapies for esophageal and gastroesophageal junction cancer. Cochrane Database of Systematic Reviews 2017 (2017).
6. Al-Batran, S.-E. & Ajani, J. A. Impact of Chemotherapy on Quality of Life in Patients With Metastatic Esophagogastric Cancer. Wiley Online Library 116, 2511–2518 (2010).
7. Moher, D., Liberati, A., Tetzlaff, J., *et al.* Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Medicine 6 (2009).
8. Efficace, F., Bottomley, A., Osoba, D., *et al.* Beyond the development of health-related quality-of-life (HRQOL) measures: A checklist for evaluating HRQOL outcomes in cancer clinical trials - Does HRQOL evaluation in prostate cancer research inform clinical decision making? Journal of Clinical Oncology 21, 3502–3511 (2003).
9. Ter Veer, E., van Kleef, J. J., Sprangers, M. A., *et al.* Reporting of health-related quality of life in randomized controlled trials involving palliative systemic therapy for esophagogastric cancer: a systematic review. Gastric Cancer 21, 183–195 (2018).
10. Guyot, P., Ades, A. E., Ouwens, M. J. & Welton, N. J. Enhanced secondary analysis of survival data: Reconstructing the data from published Kaplan-Meier survival curves. BMC Medical Research Methodology 12 (2012).
11. Osoba, D., Bezjak, A., Brundage, M., *et al.* Analysis and interpretation of health-related quality-of-life data from clinical trials: Basic approach of the National Cancer Institute of Canada Clinical Trials Group. European Journal of Cancer 41, 280–287 (2005).
12. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. Tech. rep. (2010).
13. Ter Veer, E., Mohammad, N. H., Van Valkenhoef, G., *et al.* The Efficacy and Safety of First-line Chemotherapy in Advanced Esophagogastric Cancer: A Network Meta-analysis. Journal of the National Cancer Institute 108 (2016).
14. Ajani, J. A., Moiseyenko, V. M., Tjulandin, S., *et al.* Quality of life with docetaxel plus cisplatin and fluorouracil compared with cisplatin and fluorouracil from a phase III trial for advanced gastric or gastroesophageal adenocarcinoma: the V-325 Study Group. Journal of Clinical Oncology 25, 3210–3216 (2007).
15. Curran, D., Pozzo, C., Zaluski, J., *et al.* Quality of life of palliative chemotherapy naive patients with advanced adenocarcinoma of the stomach or esophagogastric junction treated with irinotecan combined with 5-fluorouracil and folinic acid: Results of a randomised phase III trial. Quality of Life Research 18, 853–861 (2009).
16. Gubanski, M., Glimelius, B. & Lind, P. A. Quality of life in patients with advanced gastric cancer sequentially treated with docetaxel and irinotecan with 5-fluorouracil and folinic acid (leucovin). Medical Oncology 31 (2014).

17. Kripp, M., Al-Batran, S. E., Rosowski, J., *et al.* Quality of life of older adult patients receiving docetaxel-based chemotherapy triplets for esophagogastric adenocarcinoma: A randomized study of the Arbeitsgemeinschaft Internistische Onkologie (AIO). Gastric Cancer 17, 181–187 (2014).
18. Bodoky, G., Scheulen, M. E., Rivera, F., *et al.* Clinical Benefit and Health-Related Quality of Life Assessment in Patients Treated with Cisplatin/S-1 Versus Cisplatin/5-FU: Secondary End Point Results From the First-Line Advanced Gastric Cancer Study (FLAGS). Journal of Gastrointestinal Cancer 46, 109–117 (2015).
19. Nuemi, G., Devilliers, H., Le Malicot, K., *et al.* Construction of quality of life change patterns: Example in oncology in a phase III therapeutic trial (FFCD 0307). Health and Quality of Life Outcomes 13 (2015).
20. Al-Batran, S. E., Van Cutsem, E., Oh, S. C., *et al.* Quality-of-life and performance status results from the phase III RAINBOWstudy of ramucirumab plus paclitaxel versus placebo plus paclitaxel in patients with previously treated gastric or gastroesophageal junction adenocarcinoma. Annals of Oncology 27, 673–679 (2016).
21. Martin, A. J., Gibbs, E., Sjoquist, K., *et al.* Health-related quality of life associated with regorafenib treatment in refractory advanced gastric adenocarcinoma. Gastric Cancer 21, 473–480 (2018).
22. Satoh, T., Bang, Y.-J., Gotovkin, E. A., *et al.* Quality of Life in the Trastuzumab for Gastric Cancer Trial. The Oncologist 19, 712–719 (2014).
23. Ajani, J. A., Rodriguez, W., Bodoky, G., *et al.* Multicenter phase III comparison of cisplatin/S-1 with cisplatin/infusional fluorouracil in advanced gastric or gastroesophageal adenocarcinoma study: The FLAGS trial. Journal of Clinical Oncology 28, 1547–1553 (2010).
24. Al-Batran, S. E., Pauligk, C., Homann, N., *et al.* The feasibility of triple-drug chemotherapy combination in older adult patients with oesophagogastric cancer: A randomised trial of the Arbeitsgemeinschaft Internistische Onkologie (FLOT65+). European Journal of Cancer 49, 835–842 (2013).
25. Bang, Y. J., Van Cutsem, E., Feyereislova, A., *et al.* Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): A phase 3, open-label, randomised controlled trial. The Lancet 376, 687–697 (2010).
26. Bouché, O., Raoul, J. L., Bonnetain, F., *et al.* Randomized multicenter phase II trial of a biweekly regimen of fluorouracil and leucovorin (LV5FU2), LV5FU2 plus cisplatin, or LV5FU2 plus irinotecan in patients with previously untreated metastatic gastric cancer: A Fédération Francophone de Cancérologie. Journal of Clinical Oncology 22, 4319–4328 (2004).
27. Bramhall, S. R., Hallissey, M. T., Whiting, J., *et al.* Marimastat as maintenance therapy for patients with advanced gastric cancer: A randomised trial. British Journal of Cancer 86, 1864–1870 (2002).
28. Cunningham, D., Starling, N., Rao, S., *et al.* Capecitabine and Oxaliplatin for Advanced Esophagogastric Cancer. New England Journal of Medicine 358, 36–46 (2008).
29. Dank, M., Zaluski, J., Barone, C., *et al.* Randomized phase III study comparing irinotecan combined with 5-fluorouracil and folinic acid to cisplatin combined with 5-fluorouracil in chemotherapy naive patients with advanced adenocarcinoma of the stomach or esophagogastric junction. Annals of Oncology 19, 1450–1457 (2008).
30. Gubanski, M., Johnsson, A., Fernebro, E., *et al.* Randomized phase II study of sequential docetaxel and irinotecan with 5-fluorouracil/folinic acid (leucovorin) in patients with advanced gastric cancer: The GATAC trial. Gastric Cancer 13, 155–161 (2010).

31. Guimbaud, R., Louvet, C., Ries, P., *et al.* Prospective, randomized, multicenter, phase III study of fluorouracil, leucovorin, and irinotecan versus epirubicin, cisplatin, and capecitabine in advanced gastric adenocarcinoma: A French intergroup (Fédération Francophone de Cancérologie Digestive, Féd. Journal of Clinical Oncology 32, 3520–3526 (2014).
32. Hall, P. S., Lord, S. R., Collinson, M., *et al.* A randomised phase II trial and feasibility study of palliative chemotherapy in frail or elderly patients with advanced gastroesophageal cancer (321GO). British Journal of Cancer 116, 472–478 (2017).
33. Hecht, J. R., Bang, Y. J., Qin, S. K., *et al.* Lapatinib in combination with capecitabine plus oxaliplatin in human epidermal growth factor receptor 2-positive advanced or metastatic gastric, esophageal, or gastroesophageal adenocarcinoma: TRIO-013/LOGiC - A randomized phase III trial. Journal of Clinical Oncology 34, 443–451 (2016).
34. Hwang, I. G., Ji, J. H., Kang, J. H., *et al.* A multi-center, open-label, randomized phase III trial of first-line chemotherapy with capecitabine monotherapy versus capecitabine plus oxaliplatin in elderly patients with advanced gastric cancer. Journal of Geriatric Oncology 8, 170–175 (2017).
35. Kim, G. M., Jeung, H. C., Rha, S. Y., *et al.* A randomized phase II trial of S-1-oxaliplatin versus capecitabine- oxaliplatin in advanced gastric cancer. European Journal of Cancer 48, 518–526 (2012).
36. Kim, M. J., Kong, S. Y., Nam, B. H., *et al.* A randomized phase II study of S-1 versus capecitabine as first-line chemotherapy in elderly metastatic gastric cancer patients with or without poor performance status: Clinical and pharmacogenetic results. Pharmacogenetics and Genomics 28, 23–30 (2018).
37. Lu, Z., Zhang, X., Liu, W., *et al.* A multicenter, randomized trial comparing efficacy and safety of paclitaxel/capecitabine and cisplatin/capecitabine in advanced gastric cancer. Gastric Cancer 21, 782–791 (2018).
38. Ohtsu, A., Shah, M. A., Van Cutsem, E., *et al.* Bevacizumab in combination with chemotherapy as first-line therapy in advanced gastric cancer: A randomized, double-blind, placebo-controlled phase III study. Journal of Clinical Oncology 29, 3968–3976 (2011).
39. Se, H. P., Woon, K. L., Chung, M., *et al.* Paclitaxel versus docetaxel for advanced gastric cancer: A randomized phase II trial in combination with infusional 5-fluorouracil. Anti-Cancer Drugs 17, 225–229 (2006).
40. Park, S. H., Nam, E., Park, J., *et al.* Randomized phase II study of irinotecan, leucovorin and 5-fluorouracil (ILF) versus cisplatin plus ILF (PILF) combination chemotherapy for advanced gastric cancer. Annals of Oncology 19, 729–733 (2008).
41. Park, S. R., Kim, M. J., Nam, B. H., *et al.* A randomised phase II study of continuous versus stop-and-go S-1 plus oxaliplatin following disease stabilisation in first-line chemotherapy in patients with metastatic gastric cancer. European Journal of Cancer 83, 32–42 (2017).
42. Rao, S., Starling, N., Cunningham, D., *et al.* Matuzumab plus epirubicin, cisplatin and capecitabine (ECX) compared with epirubicin, cisplatin and capecitabine alone as first-line treatment in patients with advanced oesophago-gastric cancer: A randomised, multicentre open-label phase II study. Annals of Oncology 21, 2213–2219 (2010).
43. Ross, P., Nicolson, M., Cunningham, D., *et al.* Prospective randomized trial comparing mitomycin, cisplatin, and protracted venous-infusion fluorouracil (PVI 5-FU) with epirubicin, cisplatin, and PVI 5-FU in advanced esophagogastric cancer. Journal of Clinical Oncology 20, 1996–2004 (2002).
44. Roth, A. D., Fazio, N., Stupp, R., *et al.* Docetaxel, cisplatin, and fluorouracil; docetaxel and cisplatin; and epirubicin, cisplatin, and fluorouracil as systemic treatment for ad-

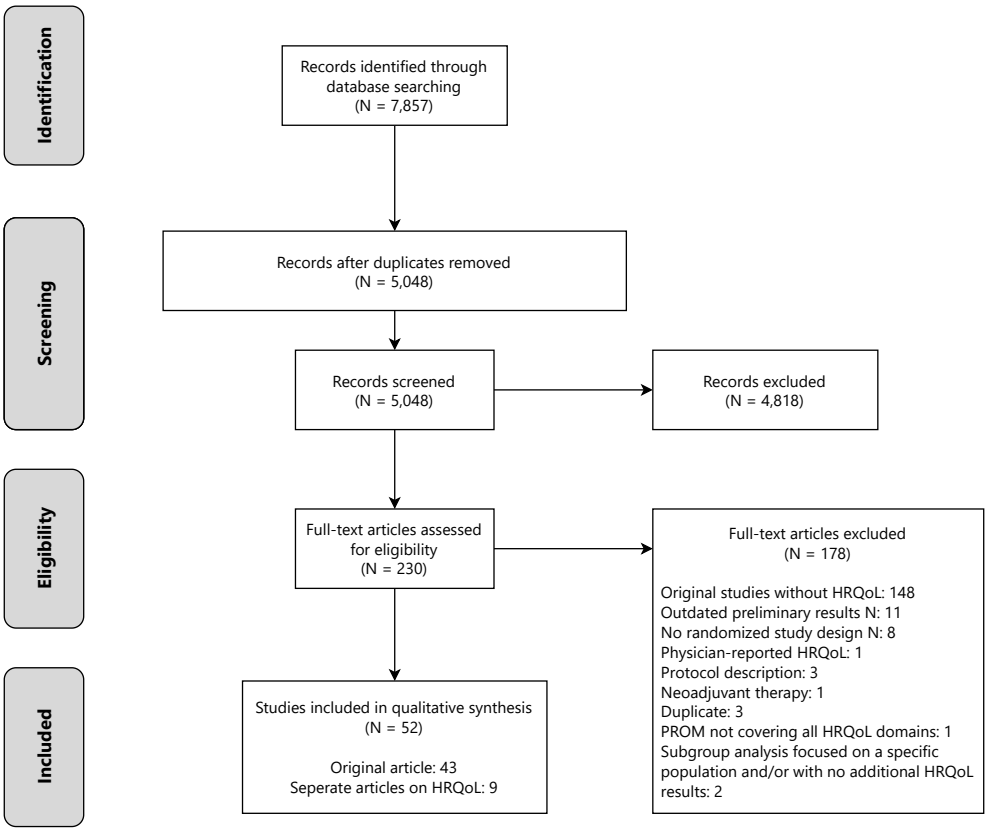
vanced gastric carcinoma: A randomized phase II trial of the Swiss group for clinical cancer research. Journal of Clinical Oncology 25, 3217–3223 (2007).

45. Ryu, M. H., Baba, E., Lee, K. H., *et al.* Comparison of two different S-1 plus cisplatin dosing schedules as first-line chemotherapy for metastatic and/or recurrent gastric cancer: A multicenter, randomized phase III trial (SOS). Annals of Oncology 26, 2097–2101 (2015).
46. Sadighi, S., Mohagheghi, M. A., Montazeri, A. & Sadighi, Z. Quality of life in patients with advanced gastric cancer: A randomized trial comparing docetaxel, cisplatin, 5-FU (TCF) with epirubicin, cisplatin, 5-FU (ECF). BMC Cancer 6 (2006).
47. Tebbutt, N. C., Norman, A., Cunningham, D., *et al.* A muticentre, randomised phase III trial comparing protracted venous infusion (PVI) 5-fluorouracil (5-FU) with PVI 5-FU plus mitomycin C in patients with inoperable oesophago-gastric cancer. Annals of Oncology 13, 1568–1575 (2002).
48. Tebbutt, N. C., Cummins, M. M., Sourjina, T., *et al.* Randomised, non-comparative phase II study of weekly docetaxel with cisplatin and 5-fluorouracil or with capecitabine in oesophagogastric cancer: The AGITG ATTAX trial. British Journal of Cancer 102, 475–481 (2010).
49. Tebbutt, N. C., Price, T. J., Ferraro, D. A., *et al.* Panitumumab added to docetaxel, cisplatin and fluoropyrimidine in oesophagogastric cancer: ATTAX3 phase II trial. British Journal of Cancer 114, 505–509 (2016).
50. Cutsem, E. V., Moiseyenko, V. M., Tjulandin, S., *et al.* Phase III Study of Docetaxel and Cisplatin Plus Fluorouracil Compared With Cisplatin and Fluorouracil As First-Line Therapy for Advanced Gastric Cancer: A Report of the V325 Study Group. Tech. rep. (2006).
51. Webb, A., Cunningham, D., Scarffe, J. H., *et al.* Randomized trial comparing epirubicin, cisplatin, and fluorouracil versus fluorouracil, doxorubicin, and methotrexate in advanced esophagogastric cancer. Journal of Clinical Oncology 15, 261–267 (1997).
52. Yoshino, S., Nishikawa, K., Morita, S., *et al.* Randomised phase III study of S-1 alone versus S-1 plus lentinan for unresectable or recurrent gastric cancer (JFMC36-0701). European Journal of Cancer 65, 164–171 (2016).
53. Bang, Y. J., Xu, R. H., Chin, K., *et al.* Olaparib in combination with paclitaxel in patients with advanced gastric cancer who have progressed following first-line therapy (GOLD): a double-blind, randomised, placebo-controlled, phase 3 trial. The Lancet Oncology 18, 1637–1651 (2017).
54. Dutton, S. J., Ferry, D. R., Blazeby, J. M., *et al.* Gefitinib for oesophageal cancer progressing after chemotherapy (COG): A phase 3, multicentre, double-blind, placebo-controlled randomised trial. The Lancet Oncology 15, 894–904 (2014).
55. Ford, H. E., Marshall, A., Bridgewater, J. A., *et al.* Docetaxel versus active symptom control for refractory oesophagogastric adenocarcinoma (COUGAR-02): An open-label, phase 3 randomised controlled trial. The Lancet Oncology 15, 78–86 (2014).
56. Fuchs, C. S., Tomasek, J., Yong, C. J., *et al.* Ramucirumab monotherapy for previously treated advanced gastric or gastro-oesophageal junction adenocarcinoma (REGARD): An international, randomised, multicentre, placebo-controlled, phase 3 trial. The Lancet 383, 31–39 (2014).
57. Li, J., Qin, S., Xu, J., *et al.* Apatinib for chemotherapy-refractory advanced metastatic gastric cancer: Results from a randomized, placebo-controlled, parallel-Arm, phase II trial. Journal of Clinical Oncology 31, 3219–3225 (2013).
58. Li, J., Qin, S., Xu, J., *et al.* Randomized, double-blind, placebo-controlled phase III trial of apatinib in patients with chemotherapy-refractory advanced or metastatic adenocar-

cinoma of the stomach or gastroesophageal junction. Journal of Clinical Oncology 34, 1448–1454 (2016).

59. Ohtsu, A., Ajani, J. A., Bai, Y. X., *et al.* Everolimus for previously treated advanced gastric cancer: Results of the randomized, double-blind, phase III GRANITE-1 study. Journal of Clinical Oncology 31, 3935–3943 (2013).
60. Pavlakis, N., Sjoquist, K. M., Martin, A. J., *et al.* Regorafenib for the treatment of advanced gastric cancer (INTEGRATE): A multinational placebo-controlled phase II Trial. Journal of Clinical Oncology 34, 2728–2735 (2016).
61. Shitara, K., Takashima, A., Fujitani, K., *et al.* Nab-paclitaxel versus solvent-based paclitaxel in patients with previously treated advanced gastric cancer (ABSOLUTE): an open-label, randomised, non-inferiority, phase 3 trial. The Lancet Gastroenterology and Hepatology 2, 277–287 (2017).
62. Thuss-Patience, P. C., Shah, M. A., Ohtsu, A., *et al.* Trastuzumab emtansine versus taxane use for previously treated HER2-positive locally advanced or metastatic gastric or gastro-oesophageal junction adenocarcinoma (GATSBY): an international randomised, open-label, adaptive, phase 2/3 study. The Lancet Oncology 18, 640–653 (2017).
63. Wilke, H., Muro, K., Van Cutsem, E., *et al.* Ramucirumab plus paclitaxel versus placebo plus paclitaxel in patients with previously treated advanced gastric or gastro-oesophageal junction adenocarcinoma (RAINBOW): A double-blind, randomised phase 3 trial. The Lancet Oncology 15, 1224–1235 (2014).
64. Lee, K.-W., Kim, B. J., Kim, M.-J., *et al.* A multicenter randomized phase II study of docetaxel vs. docetaxel plus cisplatin vs. docetaxel plus S-1 as second-line chemotherapy in metastatic gastric cancer patients who had progressed after cisplatin plus either S-1 or capecitabine. Cancer research and treatment: official journal of Korean Cancer Association 49, 706 (2017).
65. Scott, N. W., Fayers, P. M., Aaronson, N. K., *et al.* EORTC QLQ-c30 Reference Values. Brussels, Belgium: EORTC Quality of Life. `https://www.eortc.org/app/uploads/sites/2/2018/02/reference_values_manual2008.pdf` (2008).
66. Schuurhuizen, C. S., Braamse, A. M., Konings, I. R., *et al.* Does severe toxicity affect global quality of life in patients with metastatic colorectal cancer during palliative systemic treatment? A systematic review. Annals of Oncology 28, 478–486 (2017).
67. Cocks, K., King, M. T., Velikova, G., *et al.* Evidence-based guidelines for interpreting change scores for the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. European Journal of Cancer 48, 1713–1721 (2012).
68. Cocks, K., King, M. T., Velikova, G., *et al.* Evidence-based guidelines for determination of sample size and interpretation of the European organisation for the research and treatment of cancer quality of life questionnaire core 30. Journal of Clinical Oncology 29, 89–96 (2011).
69. Calvert, M., Blazeby, J., Altman, D. G., *et al.* Reporting of patient-reported outcomes in randomized trials: The CONSORT PRO extension. JAMA - Journal of the American Medical Association 309, 814–822 (2013).
70. Anota, A., Barbieri, A., Savina, M., *et al.* Comparison of three longitudinal analysis models for the health-related quality of life in oncology: A simulation study. Health and Quality of Life Outcomes 12 (2014).
71. Hagerty, R. G., Butow, P. N., Ellis, P. A., *et al.* Cancer patient preferences for communication of prognosis in the metastatic setting. Journal of Clinical Oncology 22, 1721–1730 (2004).

Supplement



Supplementary Figure 1: Flowchart of the included studies.

Supplementary Figures 2–5 are available on the website of The Journal of the National Cancer Institute via: https://doi.org/10.1093/jnci/djz133.

Database	Search terms
CENTRAL	((esophag* or oesophag* or stomach or gastric or gastroesophag* or gastrooesophag*) and (neoplas* or cancer* or carcino* or adenocarcino* or tumor or tumors or tumour or tumours or malig*)):ti,ab,kw (palliat* or advanced or metasta* or irresect* or unresect* or un-resect* or non-resect* or nonresect* or inopera* or non-opera* or nonopera* or un-opera*):ti,ab,kw (chemotherap* or polytherap* or polychemotherap* or combination* or two-agent* or two-drug* or double-drug* or doublet* or three-agent* or three-drug* or triple* or multi-agent or multi-drug or active agent* or antineoplastic* or anti-neoplastic* or anticancer* or anti-cancer* or antitumor* or anti-tumor* or antitumour* or anti-tumour* or anthracyclin* or capecitabine or carboplatin* or cisplatin* or docetaxel or doxorubicin* or epirubicin* or fluoropyrimidine* or fluorouracil or 5-FU or folinic acid or irinotecan or leucovorin* or mitomycin* or organoplatin* or oteracil or oxaliplatin* or oxonic acid or paclitaxel or platin* or S-1 or taxane* or tegafur):ti,ab,kw
EMBASE	((esophag* or oesophag* or stomach or gastric or gastroesophag* or gastrooesophag*) adj5 (neoplas* or cancer* or carcino* or adenocarcino* or tumor or tumors or tumour or tumours or malig*)).ti,ab. (palliat* or advanced or metasta* or irresect* or unresect* or un-resect* or non-resect* or nonresect* or inopera* or non-opera* or nonopera* or un-opera*).ti,ab. (chemotherap* or polytherap* or polychemotherap* or combination* or two-agent* or two-drug* or double-drug* or doublet* or three-agent* or three-drug* or triple* or multi-agent or multi-drug or active agent* or antineoplastic* or anti-neoplastic* or anticancer* or anti-cancer* or antitumor* or anti-tumor* or antitumour* or anti-tumour* or anthracyclin* or capecitabine or carboplatin* or cisplatin* or docetaxel or doxorubicin* or epirubicin* or fluoropyrimidine* or fluorouracil or 5-FU or folinic acid or irinotecan or leucovorin* or mitomycin* or organoplatin* or oteracil or oxaliplatin* or oxonic acid or paclitaxel or platin* or S-1 or taxane* or tegafur).ti,ab. exp controlled clinical trial/ or randomized.ti,ab. or randomised.ti,ab. or placebo.ti,ab. or randomly.ti,ab. or trial.ti.
Medline	(”Esophageal Neoplasms”[Mesh] OR ”Stomach Neoplasms”[Mesh] OR ((esophag*[tiab] OR oesophag*[tiab] OR stomach[tiab] OR gastric[tiab] OR gastroesophag*[tiab] OR gastrooesophag*[tiab]) AND (neoplas*[tiab] OR cancer*[tiab] OR carcino*[tiab] OR adenocarcino*[tiab] OR tumor[tiab] OR tumors[tiab] OR tumour[tiab] OR tumours[tiab] OR malig*[tiab]))) AND (”Palliative Care”[Mesh] OR ”Neoplasm Metastasis”[Mesh] OR palliat*[tiab] OR advanced[tiab] OR metasta*[tiab] OR irresect*[tiab] OR unresect*[tiab] OR un-resect*[tiab] OR non-resect*[tiab] OR nonresect*[tiab] OR inopera*[tiab] OR non-opera*[tiab] OR non-opera*[tiab] OR unopera*[tiab]) AND (”Drug Therapy, Combination”[Mesh] OR ”Drug Combinations”[Mesh] OR ”Antineoplastic Agents”[Mesh] OR ”Antineoplastic Agents”[Pharmacological Action] OR ”Anthracyclines”[Mesh] OR ”Leucovorin”[Mesh] OR ”Organoplatinum Compounds”[Mesh] OR ”Oxonic Acid”[Mesh] OR ”Taxoids”[Mesh] OR chemotherap*[tiab] OR polytherap*[tiab] OR polychemotherap*[tiab] OR combination*[tiab] OR two-agent*[tiab] OR two-drug*[tiab] OR double-drug*[tiab] OR doublet*[tiab] OR three-agent*[tiab] OR three-drug*[tiab] OR triple*[tiab] OR multi-agent[tiab] OR multi-drug[tiab] OR active agent*[tiab] OR antineoplastic*[tiab] OR anti-neoplastic*[tiab] OR anticancer*[tiab] OR anti-cancer*[tiab] OR antitumor*[tiab] OR anti-tumor*[tiab] OR antitumour*[tiab] OR anti-tumour*[tiab] OR anthracyclin*[tiab] OR capecitabine[tiab] OR carboplatin*[tiab] OR cisplatin*[tiab] OR docetaxel[tiab] OR doxorubicin*[tiab] OR epirubicin*[tiab] OR fluoropyrimidine*[tiab] OR fluorouracil[tiab] OR 5-FU[tiab] OR folinic acid[tiab] OR irinotecan[tiab] OR leucovorin*[tiab] OR mitomycin*[tiab] OR organoplatin*[tiab] OR oteracil[tiab] OR oxaliplatin*[tiab] OR oxonic acid[tiab] OR paclitaxel[tiab] OR platin*[tiab] OR S-1[tiab] OR taxane*[tiab] OR tegafur[tiab] AND (randomized controlled trial[pt] OR controlled clinical trial[pt] OR randomized[tiab] OR randomised[tiab] OR placebo[tiab] OR clinical trials as topic[mesh:noexp] OR randomly[tiab] OR trial[ti]) AND (english[la] OR dutch[la])

Supplementary Table 1: Search strategy per database.

Study	Random sequene allocation	Allocation concealment	Blinding of outcome	Incomplete outcome data	Selective reporting	Other
First-line treatment						
Ajani et al., 2010[23]; Bodoky et al., 2015[18]	Unknown	Unknown	Low	Low	Low	Low
Al-Batran et al., 2013[24] – Kripp et al., 2014[17]	Low	Low	Low	Low	Low	Low
Bang et al., 2010[25]; Satoh et al., 2014[22]	Low	Low	Low	Low	Low	Low
Bouché et al., 2004[26]	Unknown	Unknown	Low	Low	Low	Low
Bramhall et al., 2002[27]	Low	Low	Low	Low	Low	Unknown
Cunningham et al., 2008[28]	Low	Low	Low	Low	Low	Low
Dank et al., 2008[29]; Curran et al., 2009[15]	Low	Low	Low	Low	Low	Low
Glimelius et al., 1997[3]	Unknown	Unknown	Low	Low	Low	Low
Gubanski et al., 2010[30] – Gubanski et al., 2014[16]	Unknown	Unknown	Low	Low	Low	Low
Guimbaud et al., 2014[31] – Nuemi et al., 2015[19]	Low	Low	Low	Low	Low	Low
Hall et al., 2017[32]	Low	Unknown	Low	Low	Low	Low
Hecht et al., 2016[33]	Low	Low	Low	Low	Low	Low
Hwang et al., 2017[34]	Unknown	Unknown	Low	Low	Low	Low
Kim et al., 2012[35]	Low	Low	Low	Low	Low	Low
Kim et al., 2018[36]	Unknown	Unknown	Low	Low	Low	Unknown
Lu et al., 2018[37]	Low	Low	Low	Low	Low	Low
Ohtsu et al., 2011[38]	Unknown	Unknown	Low	Low	Low	Low
Park et al., 2006[39]	Unknown	Unknown	Low	Low	Low	Unknown
Park et al., 2007[40]	Unknown	Unknown	Low	Low	Low	Unknown
Park et al., 2017[41]	Unknown	Low	Low	Low	Low	Unknown
Rao et al., 2010[42]	Low	Low	Low	Low	Low	Low
Ross et al., 2002[43]	Low	Low	Low	Low	Low	Low
Roth et al., 2007[44]	Low	Low	Low	Low	Low	Low
Ryu et al., 2015[45]	Low	Low	Low	Low	Low	Low
Sadighi et al., 2006[46]	Unknown	Unknown	Low	Low	Low	Unknown
Tebbutt et al., 2002[47]	Low	Low	Low	Low	Low	Low
Tebbutt et al., 2010[48]	Low	Low	Low	Low	Low	Low
Tebbutt et al., 2016[49]	Low	Low	Low	Low	Low	Low
Van Cutsem et al., 2006[50]; Ajani et al., 2007[14]	Low	Low	Low	Low	Low	Low
Webb et al., 1997[51]	Low	Low	Low	Low	Low	Low
Yoshino et al., 2016[52]	Low	Unknown	Low	Low	Low	Low

Supplementary Table 2 – *continued from previous page*

Study	Random sequene allocation	Allocation concealment	Blinding of outcome	Incomplete outcome data	Selective reporting	Other
Beyond first-line treatment						
Bang et al., 2017[53]	Low	Low	Low	Low	Low	Low
Dutton et al., 2014[54]	Low	Low	Low	Low	Low	Low
Ford et al., 2014[55]	Low	Low	Low	Low	Low	Low
Fuchs et al., 2014[56]	Low	Low	Low	Low	Low	Low
Lee et al., 2017[64]	Low	Unknown	Low	Low	Low	Low
Li et al., 2013[57]	Low	Low	Low	Low	Low	Low
Li et al., 2016[58]	Low	Low	Low	Low	Low	Low
Ohtsu et al., 2013[59]	Low	Low	Low	Low	Low	Low
Pavlakis et al., 2016[60]; Martin et al., 2018[21]	Low	Low	Low	Low	Low	Low
Shitara et al., 2017[61]	Low	Low	Low	Low	Low	Low
Thuss-Patience et al., 2017[62]	Low	Low	Low	Low	Low	Low
Wilke et al., 2014[63]; Al-Batran et al., 2016[20]	Low	Low	Low	Low	Low	Low

Supplementary Table 2: Risk of bias.

Study	A priori hypothesis stated	Rationale for instrument reported	Psychometric properties	Cultural validity	HRQoL domains	Instrument administration	Baseline compliance	Timing of assessments	Handling and amount of missing data	Clinical significance addressed	HRQoL discussed	Quality	Interpretation
First-line treatment													
Ajani et al., 2010[23]; Bodoky et al., 2015[18]	+	−	+	+	+	+	+	+	+	+	+	0.91	Probably robust
Al-Batran et al., 2013[24] – Kripp et al., 2014[17]	+	−	+	+	+	−	+	+	−	+	+	0.73	Limited
Bang et al., 2010[25]; Satoh et al., 2014[22]	+	−	+	+	+	+	+	+	+	+	+	0.91	Probably robust
Bouché et al., 2004[26]	−	−	+	+	+	−	+	+	+	−	+	0.64	Limited
Bramhall et al., 2002[27]	−	−	+	+	+	−	−	+	−	−	−	0.36	Very limited
Cunningham et al., 2008[28]	−	−	+	+	+	−	+	+	−	−	−	0.45	Limited
Dank et al., 2008[29]; Curran et al., 2009[15]	+	−	+	+	+	−	+	+	+	+	+	0.82	Probably robust
Glimelius et al., 1997[3]	−	−	+	+	+	−	+	+	+	−	+	0.64	Limited
Gubanski et al., 2010[30] – Gubanski et al., 2014[16]	*	−	+	+	+	−	+	+	+	+	+	0.80	Probably robust
Guimbaud et al., 2014[31] – Nuemi et al., 2015[19]	−	−	+	+	+	−	−	+	+	+	+	0.64	Limited
Hall et al., 2017[32]	−	−	+	+	+	+	+	+	−	−	−	0.55	Limited
Hecht et al., 2016[33]	−	−	+	+	+	−	−	−	−	−	−	0.27	Very limited
Hwang et al., 2017[34]	−	−	+	+	+	−	+	+	−	+	−	0.55	Limited
Kim et al., 2012[35]	+	−	+	+	+	−	+	+	−	−	−	0.55	Limited
Kim et al., 2018[36]	*	−	+	+	+	−	+	+	−	−	+	0.60	Limited
Lu et al., 2018[37]	+	−	+	+	+	−	−	+	−	−	−	0.45	Limited
Ohtsu et al., 2011[38]	−	−	+	+	+	−	−	+	−	−	−	0.36	Very limited
Park et al., 2006[39]	−	−	+	+	+	−	+	+	−	+	+	0.64	Limited
Park et al., 2007[40]	−	−	+	+	+	+	+	+	−	−	+	0.64	Limited
Park et al., 2017[41]	−	−	+	+	+	−	−	+	−	−	+	0.55	Limited
Rao et al., 2010[42]	−	−	+	+	+	−	+	+	−	−	−	0.45	Limited
Ross et al., 2002[43]	−	−	+	+	+	−	+	+	−	−	+	0.55	Limited
Roth et al., 2007[44]	−	−	+	+	+	−	+	+	−	+	+	0.64	Limited
Ryu et al., 2015[45]	−	−	+	+	+	−	+	+	−	−	−	0.45	Limited
Sadighi et al., 2006[46]	−	−	+	+	+	−	+	+	+	+	+	0.73	Probably robust

Supplementary Table 3 – *continued from previous page*

Study	A priori hypothesis stated	Rationale for instrument reported	Psyhometric properties	Cultural validity	HRQoL domains	Instrument administration	Baseline compliance	Timing of assessments	Handling and amount of missing data	Clinical significance addressed	HRQoL discussed	Quality	Interpretation
Tebbutt et al., 2002[47]	–	–	+	+	+	+	–	+	–	–	–	0.45	Limited
Tebbutt et al., 2010[48]	–	–	+	+	+	–	+	+	–	+	–	0.55	Limited
Tebbutt et al., 2016[49]	–	–	+	+	+	–	+	+	–	+	–	0.55	Limited
Van Cutsem et al., 2006[50]; Ajani et al., 2007[14]	+	–	+	+	+	+	+	+	+	+	+	0.91	Probably robust
Webb et al., 1997[51]	–	–	+	+	+	–	+	+	+	–	+	0.64	Limited
Yoshino et al., 2016[52]	–	–	+	+	+	–	+	+	–	–	+	0.55	Limited
Beyond first-line treatment													
Bang et al., 2017[53]	–	–	+	+	+	–	–	–	–	+	–	0.36	Limited
Dutton et al., 2014[54]	*	–	+	+	+	+	+	+	+	+	+	0.90	Probably robust
Ford et al., 2014[55]	+	–	+	+	+	+	+	+	+	+	+	0.91	Probably robust
Fuchs et al., 2014[56]	–	–	+	+	+	–	+	+	+	+	+	0.73	Probably robust
Lee et al., 2017[64]	–	–	+	+	+	–	–	+	–	+	–	0.45	Limited
Li et al., 2013[57]	–	–	+	+	+	–	–	+	–	–	–	0.36	Very limited
Li et al., 2016[58]	–	–	+	+	+	–	+	+	–	–	–	0.45	Limited
Ohtsu et al., 2013[59]	+	–	+	+	+	–	+	+	–	+	–	0.64	Limited
Pavlakis et al., 2016[60]; Martin et al., 2018[21]	*	–	+	+	+	–	+	+	+	+	+	0.80	Probably robust
Shitara et al., 2017[61]	–	–	+	+	+	–	+	+	–	–	+	0.55	Limited
Thuss-Patience et al., 2017[62]	–	–	+	+	+	–	–	–	–	–	–	0.27	Very limited
Wilke et al., 2014[63]; Al-Batran et al., 2016[20]	+	–	+	+	+	–	+	+	+	+	+	0.82	Probably robust

Supplementary Table 3: HRQoL quality assessment.

A study classified as 'probably robust' has fulfilled at least the following three criteria: psychometric properties, baseline compliance, and handling and amount of missing data. +: the criterion of interest is described in the study report; –: the criterion of interest is not (fully) described in the study report; *: the analysis had an exploratory nature.

Chapter 8

Health-related quality of life in curatively-treated patients with oesophageal or gastric cancer

A systematic review and meta-analysis

Héctor G. van den Boorn*, Charlotte I. Stroes*, Aeilko H. Zwinderman, Wietse J. Eshuis, Maarten C.C.M. Hulshof, Faridi S. van Etten-Jamaludin, Mirjam A.G. Sprangers and Hanneke W.M. van Laarhoven

Based on:

H.G. van den Boorn*, C.I. Stroes*, A.H. Zwinderman, *et al.* Health-related quality of life in curatively-treated patients with esophageal or gastric cancer: a systematic review and meta-analysis. Critical Reviews in Oncology/ Hematology 154 (2020).

* *These authors contributed equally*

Abstract

Surgery and chemoradiotherapy can potentially cure oesophageal and gastric cancer patients, although they may impact health-related quality of life (HRQoL). We aim to systemically review and meta-analyse literature to determine the effect of curative treatments on HRQoL in oesophageal and gastric cancer. A systematic search was performed identifying studies assessing HRQoL. Meta-analyses were performed on baseline and subsequent timepoints.

From the 6,067 articles retrieved, 49 studies were included (61% low quality). Meta-analyses showed short-term HRQoL differences between oesophageal cancer patients receiving definitive chemoradiotherapy (dCRT), neoadjuvant chemo(radio)therapy (nC(R)T), or surgery alone ($p < 0.001$), with better HRQoL with nC(R)T and surgery compared to dCRT. Over the course of 12 months, no HRQoL difference was identified between treatments in oesophageal cancer ($p = 0.633$). Oesophagectomy, but not gastrectomy, resulted in a clinically relevant decline in HRQoL. No long-term HRQoL differences were identified between curative treatments in oesophageal and gastric cancer. More high-quality HRQoL studies are warranted.

Introduction

Both gastric and oesophageal cancer are highly lethal diseases, ranking third and sixth as leading causes of cancer mortality worldwide, respectively.[1] Despite intensive multimodality therapy, recurrence rates are high[2,3] and the five-year overall survival rate is poor, rarely exceeding 40%.[3–6]

Large differences exist in clinical practice for both tumours throughout the world, due to local preferences and regional variation in tumour characteristics.[7] Currently, the preferred curative treatment options for gastric cancer are gastrectomy with perioperative chemotherapy in Europe[8,9], adjuvant chemotherapy in Asia[10], or adjuvant chemoradiotherapy in the United States[11]. Neoadjuvant chemo(radio)therapy (nC(R)T) followed by oesophagectomy is the preferred treatment strategy for oesophageal cancer in the United States and Europe[3,12,13], whereas nC(R)T or adjuvant chemotherapy are both applied in Asia[14].

Although patients with oesophageal and gastric cancer may benefit from potentially-curative treatments in terms of survival, treatment could have a profound impact on patients' health-related quality of life (HRQoL).[15–18] It has already been demonstrated that major surgery, such as oesophagectomy and gastrectomy, could cause deterioration of HRQoL.[15–17,19–22] Health-related quality of life, both before and after treatment, is of great importance to cancer patients. Thus, HRQoL is increasingly recognised as an important component in the process of decision making for both physicians and patients.[23] Moreover, it has also been shown that HRQoL data potentially hold prognostic value, in addition to clinical measures.[24]

Here, we aim to present an overview of published HRQoL studies of both patients with oesophageal and gastric cancer who are treated with curative intent. We performed meta-analyses to analyse: (I) the short-term impact of treatment on HRQoL in oesophageal cancer patients receiving nC(R)T, definitive chemoradiotherapy (dCRT), or surgery alone; (II) the impact of dCRT or surgery in oesophageal cancer patients up to one year post-treatment; (III) the impact of nC(R)T followed by surgery or surgery alone in oesophageal cancer patients over the first 12 months; (IV) the impact of surgery in gastric cancer patients over the first 12 months.

Methods

Search strategy

We performed a systematic review in line with the guidelines of Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA).[25] We systematically searched Medline, EMBASE, Central, PsychINFO, and CINAHL through October 2019 to identify eligible articles assessing the HRQoL in patients with oesophageal cancer or gastric cancer treated with curative intent. The search included terms for tumour location (oesophageal or gastric cancer), treatment modalities (surgery, radiotherapy or chemotherapy) and HRQoL (see Supplementary Table 1 for full search). Two authors (HvdB and CS) independently screened all potentially eligible studies using titles and abstracts, and analysed the full-text selection for eligibility. Disagreements were resolved by discussion, or with a third arbiter (HvL).

Eligibility criteria

We considered studies eligible if studies included patients with gastric or oesophageal cancer undergoing potentially curative treatment, evaluated HRQoL data using validated health-related quality of life questionnaires, and reported HRQoL data at baseline prior to treatment initiation. Studies were excluded if only patients with precursor lesions or T_1 tumours were included, if both patients with oesophageal and gastric cancer were included in a study, the publication date was prior to 2000, the article was not written in English, or if non-validated or self-constructed questionnaires were used. Reviews and case-reports were also excluded.

Data extraction

Data from eligible studies were extracted independently by two authors (HvdB and CS) using a pre-defined extraction sheet, obtaining information on key inclusion criteria, baseline characteristics, treatment and HRQoL assessment tools. From all available subscales at baseline and follow-up, we obtained the HRQoL mean or median scores, standard deviations, and sample sizes. Studies concerning surgery alone were classified as 'untreated prior to surgery' if no more than 20% of patients received any form of pre-treatment. Otherwise, the studies were classified as 'potentially pre-treated'. Tumours of the gastro-intestinal junction were classified as gastric cancer if the Siewert classification was III; if no Siewert classification was specified, the tumour locations were classified according to the original study.[26] If the data was unclear, the corresponding author was contacted by e-mail for clarification. Plot Digitizer 2.6.8 was used to obtain values from figures.

The primary outcomes were the European Organisation for Research and Treatment of Cancer (EORTC) QLQ-C30 Global Health Status (GHS), an overall measure for health-related quality of life, and physical functioning. Other subscales served as secondary outcomes. The EORTC OES18 and OES24 questionnaires were combined during our analyses. Analyses were performed on all available questionnaires and subscales, if data from at least three studies per treatment group were available.

Study quality and risk of bias assessment

The Cochrane risk of bias tool was used to assess randomised trials.[27] Non-randomised trials were also scored with the Cochrane risk of bias tool on incomplete outcome data, selective outcome reporting and other bias. Furthermore, all trials were critically assessed on HRQoL outcome reporting using both the modified version of the 'minimum standard checklist for evaluating HRQoL outcomes' (MSCEH)[28] and the 'Newcastle-Ottawa Quality Assessment Scale' (NOQAS)[29] as has been reported previously[30]. All domains were scored as 'low risk', 'high risk', 'unclear risk', or 'not applicable' by two authors (HvdB, CS).

Studies were identified as 'high-quality' if two mandatory items and at least three out of four additional criteria were assessed as 'low risk'. The mandatory items were sample size of at least 26 at baseline (NOQAS) and a robust HRQoL reporting, which we defined as reporting of mean/median scores, sample sizes and confidence interval or its equivalent. The additional criteria were: description of treatment (NOQAS), incomplete reporting (Cochrane), selective reporting (Cochrane), and coverage domain (MSCEH), defined as the use of a cancer-site (oesophageal or gastric) specific questionnaire.

Statistical analyses

Meta-analyses on HRQoL scales were performed on the extracted data in four different analyses. First, the short-term effects at the start of treatments on HRQoL in the first four months were studied in oesophageal cancer patients receiving nC(R)T, surgery alone or dCRT. We analysed the nC(R)T data points prior to surgery, as we were interested in the specific effect of nCRT on HRQoL without including the effect of surgery on HRQoL. This is currently also being investigated as curative treatment option with active surveillance.[31] Second, dCRT and surgery (both untreated and potentially pre-treated) for oesophageal cancer were compared, with a follow-up of 12 months from baseline. Third, oesophageal cancer patients who received nC(R)T combined with surgery or surgery alone were compared with a follow-up of 12 months start-

ing from surgery. The interval between nC(R)T and surgery was extracted from the applicable studies. If the latter was not reported, an interval of 3.3 months was imputed, based on the median interval of available data. Finally, for gastric cancer the impact of surgery was analysed, starting at baseline with a follow-up of 12 months. Furthermore, as treatments and survival outcomes are known to be different between Asia and the Western world, we also performed analyses on geographical differences in HRQoL outcomes between studies conducted in Asia and the Western world.

All analyses were performed in the R studio environment (R version 3.6.1)[32] with the *metafor*[33] package, version 2.1-0. A logit transformation was applied to all subscales with a range of 0–100, in order to preserve this range during analyses. As most studies did not report standard errors on the reported HRQoL data, the errors were estimated with the mean scores and sample size, using the *metafor* package. All data were analysed with mixed-effects linear models, with the individual studies added as random effects. The fixed effects consisted of main effects and an interaction between treatment arm (surgery, nC(R)T and dCRT) and time. A quadratic spline transformation was used to capture non-linear effects in time in the second, third and fourth meta-analyses. It was, however, not employed in the first analysis because of the short follow-up. Various options of the spline parameters (i.e. the knot points) were assessed and selected according to the minimum value of the Akaike Information Criterion.[34]

For each analysis, the difference between treatment arms was assessed with an omnibus chi-squared test and, in case of significance, one-sided post-hoc tests were performed to identify differences between individual treatment arms. The post-hoc tests were corrected for multiple comparisons with the Holm method.[35] The difference in subscale score between baseline and maximum follow-up was also analysed, as well as the difference between treatment arms at maximum follow-up. A significance level of $p = 0.05$ was set for all statistical analyses. A clinically significant difference or change was set at 10 points.

Results

Search

From the 6,067 original articles retrieved, 49 studies were included (Figure 1). Two articles[17,36] described the same study population and were thus considered as a single study. The main reasons for exclusion following full-text screening were absence of baseline data (N = 17), absence of HRQoL scores or time references (N = 9), no full-text availability (N = 7), or inclusion of only $\leq T_1$ tumours (N = 7).

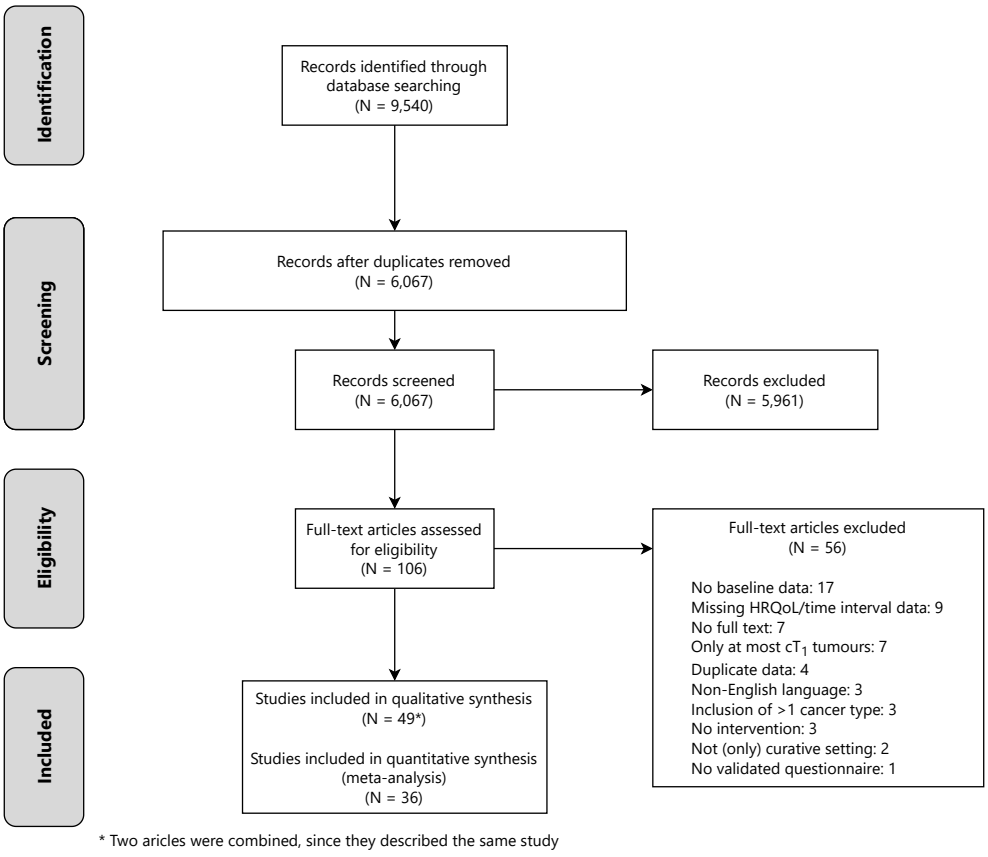


Figure 1: Flow diagram of included studies according to the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA).

Identified articles included patients between 1993–2018. The majority of included studies (N = 40) included patients with oesophageal cancer[17,18,36–73], and nine articles included gastric cancer patients[74–82]. We contacted three corresponding authors for clarification of the reported data.[48,51,67] The major baseline characteristics of the included studies are listed in Table 1.

HRQoL measures

The majority of studies used more than one questionnaire (N = 32, 65%). The cancer-specific QLQ-C30 questionnaire from the EORTC[83] was used most frequently (N = 37, 76%). This questionnaire consists of 30 questions, which result in the GHS subscale, five functional scales (physical, emotional, role, cognitive, and social), in which higher scores represent better HRQoL, and nine symptom scales (fatigue, pain, nausea/vomiting, dyspnoea, insomnia, appetite loss, constipation, diarrhoea, financial difficulties), in which higher scores represent a higher symptom burden. Other frequently used questionnaires were the oesophageal cancer-specific questionnaires EORTC QLQ-OES18[70] (N = 14, 29%) and QLQ-OES24[84] (N = 5, 10%), and the gastric-cancer specific questionnaire EORTC QLQ-STO22[74] (N = 7, 14%). Other studies also applied the FACT-E[85]/FACT-HN[86] (N = 6, 12%), MOS-SF20[87] (N = 3, 6%), RSCL[88] (N = 3, 6%), Spitzer Index[89] (N = 2, 4%), POMS[90] (N = 1, 2%), BREF[91] (N = 1, 2%), Kuechler et al.[92] (N = 1, 2%), and Korenaga et al.[93] (N = 1, 2%).

Quality of studies

We considered 19 studies (39%) to be of high quality, whereas the majority of studies (N = 30, 61%) were of low quality (Figure 2; Supplementary Table 2). Twenty-three of the 49 studies demonstrated high risk of bias in one or more of the mandatory criteria (N $\geq$ 26 at baseline and/or robust HRQoL reporting).

Meta-analyses

Thirty-six out of 49 studies were included in the meta-analyses. Only studies applying EORTC questionnaires were included, as with the other questionnaires fewer than three articles per treatment group were available.

Study	Type of cancer	Study design	Inclusion years	Country	HRQoL measures	Intervention	N
Alderson 2017[57]	Oesophageal	RCT	2005–2011	UK	QLQ-C30, QLQ-OES18	Neoadjuvant Cis + [FU vs. Epi + Cap] + Surgery	897
Avery 2007[68]	Oesophageal	Cohort	2000–2004	UK	QLQ-C30, QLQ-OES18	[Definitive Cis + FU + RT] vs. [Neoadjuvant Cis + FU + RT + surgery]	132
Barbour 2008[69]	Oesophageal	Cohort	2000–2003	UK	QLQ-C30	TG vs. TTO	63
Barbour 2017[94]	Oesophageal	Cohort	1998–2011	Australia	QLQ-C30, QLQ-OES18	Thoracoscopically assisted McKeown vs. Open Ivor-Lewis	487
Blazeby 2003[70]	Oesophageal	Cohort	1998–2001	Belgium	QLQ-C30, QLQ-OES24	Surgery vs. dCRT	591
Blazeby 2004[74]	Gastric	Case-series	2001–2003	Multiple EU-countries	QLQ-STO22, QLQ-C30	TG or PG	108
Blazeby 2005[71]	Oesophageal	Cohort	2000–2003	UK	QLQ-C30, QLQ-OES18	[Neoadjuvant Cis + FU + RT + surgery] vs. [Cis + FU (+Epi) + surgery] vs. [surgery alone]	103
Bonnetain 2006[72]	Oesophageal	RCT	1993–2000	France	Spitzer Index	[nCRT + surgery] vs. dCRT	259
Brooks 2002[73]	Oesophageal	Cohort	1998–2000	USA	FACT-E, POMS	[nC(R)T + surgery] vs. surgery alone	38
Cavallin 2015[37]	Oesophageal	Cohort	2009–2014	Italy	QLQ-C30, QLQ-OES18	Ivor-Lewis vs. McKeown	109
Cense 2006[38]	Oesophageal	Cohort	1994–2000	The Netherlands	MOS-SF20, RSCL	TTO	104
Chou 2010[39]	Oesophageal	Cohort	2002–2007	Taiwan	BREF	Surgery vs. [Definitive Cis + FU + leuco]	29
De Boer 2004[40]	Oesophageal	RCT	1994–2000	The Netherlands	MOS-SF20, RSCL	THO vs. TTO	199
Egberts 2008[41]	Oesophageal	Cohort	1998–2006	Germany	QLQ-C30, Kuechler et al.[92]	TTO vs. McKeown	105
Fang 2004[42]	Oesophageal	Case-series	1999–2001	Taiwan	QLQ-C30	Definitive Cis + FU + RT	110
Garant 2019[43]	Oesophageal	Cohort	2015–2018	USA	FACT-E	Carbo + Pacli + (surgery) + [PRT vs. XRT]	125
Gillham 2008[44]	Oesophageal	Case-series	1998–2005	Ireland	QLQ-C30, QLQ-OES24	Definitive Cis + FU + RT	56
Gradauskas 2006[45]	Oesophageal	Cohort	NA	Lithuania	QLQ-C30	Ivor-Lewis vs. McKeown	49
Haj Mohammad 2016[46]	Oesophageal	Case-series + Cross-sectional	2012–2014	The Netherlands	QLQ-C30, QLQ-OES18	nCRT + surgery	76
Hauser 2015[47]	Oesophageal	Cohort	1998–2009	Germany	QLQ-C30	[nC(R)T + surgery] vs. surgery alone	131
Huang 2015[48]	Oesophageal	Case-series	2012–2013	China	QLQ-C30, QLQ-OES18	Surgery	196
Kachnic 2011[49]	Oesophageal	RCT	1995–1999	USA	FACT-HN	Definitive FU + Cis + [50.4 Gy RT vs. 64.8 Gy RT]	166
Kassam 2010[75]	Gastric	Case-series	NA	Canada	QLQ-C30	Adjuvant FU + Cis + RT	33
Kim 2012[76]	Gastric	Cohort	2005–2007	Korea	QLQ-C30, GLG-STO22	PG vs. TG	465
Kobayashi 2011[77]	Gastric	Cohort	2005–2007	Japan	QLQ-C30, QLQ-STO22	TG vs. open PG	98
Kong 2012[78]	Gastric	Case-series	2008–2011	South Korea	QLQ-C30, QLQ-STO22	Open or laparoscopy-assisted surgery	272
Li 2018[79]	Gastric	Cohort	2013–2015	China	QLQ-C30, QLQ-STO22	[PG + oesophagogastrostomy] vs. [PG + gastrojejunostomy] vs. [TG + Roux- en Y reconstruction]	43
Lv 2014[50]	Oesophageal	Cohort	2011–2013	China	QLQ-C30	Surgery vs. [definitive Cis + Doc + RT]	102
Malström 2015[51]	Oesophageal	Case-series	NA	Sweden	QLQ-C30, QLQ-OES18	TTO	79
Nafteux 2013[52]	Oesophageal	Case-series	2005–2009	Belgium	QLQ-C30, QLQ-OES18	Surgery	455
Noordman 2018[17,36]	Oesophageal	RCT	2004–2008	The Netherlands	QLQ-C30, QLQ-OES24	[Neoadjuvant Carbo + Pacli + RT + surgery] vs. surgery alone	363
Parameswaran 2010[53]	Oesophageal	Case-series	2005–2007	UK	QLQ-C30, QLQ-OES18	MIE	62
Park 2018[80]	Gastric	Case-series	2011–2014	South Korea	QLQ-C30, QLQ-STO22	Open TG or laparoscopy-assisted TG	417
Pruthi 2018[81]	Gastric	Case-series	NA	India	QLQ-C30, QLQ-STO22	Adjuvant FU + Leuco + RT	30
Ramakrishnaiah 2014[54]	Oesophageal	Cohort	2007–2008	India	QLQ-C30, QLQ-OES18	THO vs. TTO	55
Rees 2015[55]	Oesophageal	Cohort	2008–2012	UK	QLQ-C30	Definitive Cis + Cap + RT + [Cetuximab vs. no Cetuximab]	258
Reynolds 2006[56]	Oesophageal	Cohort	?1999	Ireland	QLQ-C30, QLQ-OES24	[nCRT + surgery] vs. surgery alone	202
Safieddine 2009[18]	Oesophageal	Case-series	2002–2005	Canada	FACT-E	Cis + Irinotecan + RT + surgery	53

Table 1 – *continued from previous page*

Study	Type of cancer	Study design	Inclusion years	Country	HRQoL measures	Intervention	N
Sarkaria 2019[58]	Oesophageal	Cohort	2012–2014	USA	FACT-E	Open surgery vs. MIE	170
Sugawara 2019[59]	Oesophageal	Cohort	2015–2017	Japan	QLQ-C30, QLQ-OES18	THO vs. TTO	37
Sunde 2019[60]	Oesophageal	RCT	2006–2013	Sweden	QLQ-C30, QLQ-OES24/ QLQ-OG25	Cis + FU + surgery + [RT vs. no RT]	181
Tatematsu 2013a[61]	Oesophageal	Case-series	2009–2010	Japan	QLQ-C30	[nCT vs. nCRT]+ surgery	27
Tatematsu 2013b[62]	Oesophageal	Case-series	2009–2010	Japan	QLQ-C30	Surgery	30
van der Sluis 2019[63]	Oesophageal	RCT	2012–2016	The Netherlands	QLQ-C30	Robot-assisted MIE vs. open TTO	109
Van Heijl 2010[64]	Oesophageal	Case-series	1994–2000	The Netherlands	MOS-SF20, RSCL	TTO vs. THO	199
Van Meerten 2008[65]	Oesophageal	Case-series	2001–2004	The Netherlands	QLQ-C30, QLQ-OES18	Carbo + Pacli + RT + surgery	54
Wu 2008[82]	Gastric	RCT	1993–1999	Taiwan	Spitzer Index, Korenaga et al.[93]	Surgery + [D1 lymphadenectomy vs. D3 lymphadenectomy]	214
Yamashita 2014[66]	Oesophageal	Case-series	2008–2010	Japan	FACT-E	Definitive FU + Nedaplatin + RT	80
Zeng 2012[67]	Oesophageal	RCT	2010	China	QLQ-C30, QLQ-OES18	Ivor-Lewis vs. [thoracoscopic + laparoscopic surgery] vs. left TTO	90

Table 1: Overview of major baseline characteristics of included studies.
RCT: randomised controlled trial; Cis: cisplatin; FU: fluorouracil; Epi: epirubicin; Cap: capecitabine; Doc: docetaxel; Leuco: leucovorin; Carbo: carboplatin; Pacli: paclitaxel; Irino: irinotecan; XRT: photon chemoradiotherapy; PRT: proton chemoradiotherapy; RT: radiotherapy; Gy: Gray; nCRT: neoadjuvant chemoradiotherapy; dCRT: definitive chemoradiotherapy; TG: total gastrectomy; PG: partial gastrectomy; TTO: transthoracic oesophagectomy; THO: transhiatal oesophagectomy; MIE: minimally invasive oesophagectomy. Ivor Lewis procedure: transthoracic oesophagectomy with an intrathoracic anastomosis; McKeown procedure: transthoracic oesophagectomy with a cervical anastomosis.

	N≥26 first point^	Robust reporting	Description treatment reported^	Incomplete outcome data*	Selective reporting*	Coverage domains°	Final assesment
Alderson 2017	+	+	+	+	+	+	+
Avery 2007	+	+	+	-	+	+	+
Barbour 2008	-	+	+	+	+	-	-
Barbour 2017	+	+	+	?	+	+	+
Blazeby 2003	+	-	-	+	-	+	-
Blazeby 2004	+	+	+	+	+	+	+
Blazeby 2005	-	-	+	+	-	+	-
Bonnetain 2006	+	+	+	?	+	-	-
Brooks 2002	-	+	-	+	+	+	-
Cavallin 2015	-	-	+	-	-	+	-
Cense 2006	+	-	+	-	-	-	-
Chou 2010	-	-	+	?	-	+	-
De Boer 2004	+	+	+	+	-	-	-
Egberts 2008	+	+	+	+	+	+	+
Fang 2004	+	+	+	+	+	-	+
Garant 2019	+	-	+	+	-	+	-
Gillham 2008	+	-	+	+	+	+	-
Gradauskas 2006	-	-	+	-	-	-	-
Haj Mohammad 2016	+	+	+	+	+	+	+
Hauser 2015	+	-	-	+	-	+	-
Huang 2015	+	-	+	-	?	+	-
Kachnic 2011	+	+	+	+	-	-	-
Kassam 2010	+	-	+	?	+	-	-
Kim 2012	+	+	+	?	+	+	+
Kobayashi 2011	+	?	+	-	+	+	-

	N≥26 first point^	Robust reporting	Description treatment reported^	Incomplete outcome data*	Selective reporting*	Coverage domains°	Final assesment
Kong 2012	?	+	-	-	+	+	-
Li 2018	+	+	+	?	-	+	-
Lv 2014	?	-	+	?	+	-	-
Malström 2015	+	?	-	+	+	+	-
Nafteux 2013	+	?	+	-	-	+	-
Noordman 2018	+	+	+	+	+	+	+
Parameswaran 2010	+	+	+	+	-	+	+
Park 2018	+	?	+	-	+	+	-
Pruthi 2018	+	-	+	-	?	+	-
Ramakrishnaiah 2014	-	-	+	-	+	+	-
Rees 2015	+	+	+	+	+	+	+
Reynolds 2006	+	-	-	+	-	+	-
Safieddine 2009	+	+	+	+	+	+	+
Sarkaria 2019	+	+	+	+	+	+	+
Sugawara 2019	+	+	+	+	-	+	+
Sunde 2019	+	+	+	+	+	+	+
Tatematsu 2013a	+	-	-	-	+	-	-
Tatematsu 2013b	+	-	-	-	-	-	-
van der Sluis 2019	?	-	+	-	-	+	-
Van Heijl 2010	+	+	+	+	+	-	+
Van Meerten 2008	+	+	+	+	+	+	+
Wu 2008	+	+	+	+	+	+	+
Yamashita 2014	+	+	+	+	+	+	+
Zeng 2012	?	-	+	-	-	+	-

Figure 2: Identification of the quality of included studies.
+: sufficient; –: insufficient; ?: unclear. N $\geq$ 26 and Robust reporting were mandatory to achieve 'High Quality' as final assessment. *Derived from the Cochrane risk of bias tool[27]; °derived from the adapted minimum standard checklist for evaluating HRQoL outcomes[28], as described previously[30]; ^derived from the adapted Newcastle-Ottawa Quality Assessment Scale[29] as previously described[30].

			Analysis 1		Analysis 2		Analysis 3		Analysis 4	
QLQ	Sub-scale	Treatment	Baseline	Max. FU	Baseline	Max. FU	Baseline	Max. FU	Baseline	Max. FU
C30	AP	dCRT	34.8 (24.8-46.3)	44.7 (19.4-29.9)	34.6 (26.4-43.7)	21.8 (26.4-43.7)*	–	–	–	–
C30	AP	nC(R)T + surgery	32.9 (27.5-38.9)	30.5 (24.8-46.3)	–	–	36.7 (30.1-43.8)	22.5 (29-39.3)**	–	–
C30	AP	Surgery	24.3 (19.4-29.9)	48.2 (27.5-38.9)***	25 (22.3-27.8)	22.2 (22.3-27.8)	12.3 (8.9-16.8)	21.2 (8.9-16.8)*	21.4 (14.5-30.4)	23 (14.5-30.4)
C30	CFˆ	dCRT	81.1 (76.3-85.1)	75.2 (80.8-86.2)	81.8 (76.5-86.1)	83.8 (76.5-86.1)	–	–	–	–
C30	CFˆ	nC(R)T + surgery	88.1 (84.3-91)	83 (84.3-91)	–	–	89 (81.9-93.6)	84.7 (78.8-86.1)	–	–
C30	CFˆ	Surgery	83.7 (80.8-86.2)	74.4 (76.3-85.1)***	84.1 (82.1-85.9)	82.5 (82.1-85.9)	87.3 (83.4-90.4)	81.4 (83.4-90.4)	88.1 (81.5-92.5)	80.1 (81.5-92.5)
C30	CO	nC(R)T + surgery	16.1 (12-21.4)	24.2 (8.3-15.5)	–	–	14.4 (8.3-23.8)	8.3 (17.4-32.8)	–	–
C30	CO	Surgery	11.4 (8.3-15.5)	16.9 (12-21.4)	–	–	8.7 (5.2-14)	12.2 (5.2-14)	15.5 (11.6-20.5)	14.9 (11.6-20.5)
C30	DI	dCRT	18.9 (8.8-35.9)	21.1 (6-15.8)	13.3 (6.9-24)	6.6 (6.9-24)	–	–	–	–
C30	DI	nC(R)T + surgery	5.5 (3-10)	14.4 (8.8-35.9)	–	–	6.4 (2-18.7)	18.4 (8.8-21.7)	–	–
C30	DI	surgery	9.8 (6-15.8)	30.2 (3-10)**	8.9 (6.2-12.5)	19.6 (6.2-12.5)**	5.3 (2.9-9.4)	21.4 (2.9-9.4)**	5.2 (3-8.7)	22 (3-8.7)***
C30	DY	dCRT	16.1 (7.6-30.8)	38.1 (8.6-20)	10.8 (5.9-19.1)	20.1 (5.9-19.1)	–	–	–	–
C30	DY	nC(R)T + surgery	9 (5.5-14.4)	22.7 (7.6-30.8)*	–	–	7 (2.7-17)	21.7 (15.2-30.6)	–	–
C30	DY	Surgery	13.3 (8.6-20)	34.9 (5.5-14.4)**	17.9 (14.5-21.9)	25 (14.5-21.9)*	7.9 (4.8-12.6)	24.4 (4.8-12.6)**	8.3 (4.7-14.3)	17 (4.7-14.3)
C30	EFˆ	dCRT	74 (69.1-78.4)	73.9 (65.6-72.3)	74.5 (68.9-79.3)	80.5 (68.9-79.3)	–	–	–	–
C30	EFˆ	nC(R)T + surgery	74.2 (69.4-78.5)	72.1 (69.1-78.4)	–	–	73.7 (63.6-81.8)	80.7 (66.9-76.4)	–	–
C30	EFˆ	Surgery	69 (65.6-72.3)	72.4 (69.4-78.5)	70.3 (67.4-73)	78.2 (67.4-73)***	66.6 (60-72.5)	76.4 (60-72.5)*	74.2 (65.3-81.5)	81.1 (65.3-81.5)
C30	FA	dCRT	38.1 (29.7-47.2)	54.4 (23.9-36.1)*	33 (27.2-39.5)	36.4 (27.2-39.5)	–	–	–	–
C30	FA	nC(R)T + surgery	24.8 (18.7-32.2)	43.1 (29.7-47.2)**	–	–	15.7 (10.5-22.8)	29.9 (33.3-45.2)*	–	–
C30	FA	Surgery	29.6 (23.9-36.1)	55.2 (18.7-32.2)***	28.4 (25.7-31.3)	33.2 (25.7-31.3)	20.4 (15.8-25.8)	34.2 (15.8-25.8)**	–	–
C30	FI	nC(R)T + surgery	11.2 (4.4-25.7)	14.5 (2.2-15.6)	–	–	18.1 (5.7-44.7)	12.3 (6.1-30.4)	–	–
C30	FI	Surgery	6.1 (2.2-15.6)	10.2 (4.4-25.7)	–	–	4.9 (1.7-13.8)	12.5 (1.7-13.8)	21.7 (16.9-27.3)	18.9 (16.9-27.3)
C30	GHSˆ	dCRT	56.6 (50.3-62.6)	49.7 (55.4-64)	56 (49.7-62.1)	69.7 (49.7-62.1)**	–	–	–	–
C30	GHSˆ	nC(R)T + surgery	69.3 (64.8-73.5)	58.7 (50.3-62.6)**	–	–	74.3 (68.9-79)	67 (57.3-65)	–	–
C30	GHSˆ	Surgery	59.8 (55.4-64)	52.6 (64.8-73.5)	61.6 (58.8-64.4)	65.1 (58.8-64.4)	67.6 (62.1-72.6)	65.5 (62.1-72.6)	68.4 (60.8-75.2)	68.1 (60.8-75.2)
C30	SL	nC(R)T + surgery	26 (20.8-32)	27.3 (19.7-28.4)	–	–	29.2 (21.8-38)	17.6 (22.2-35)*	–	–
C30	SL	Surgery	23.7 (19.7-28.4)	32.6 (20.8-32)*	–	–	18.5 (13.8-24.3)	26.1 (13.8-24.3)	22.8 (16.3-30.8)	19.2 (16.3-30.8)
C30	NV	dCRT	17.5 (12.7-23.7)	19 (11.9-19)	16.9 (12.7-22.2)	12.2 (12.7-22.2)	–	–	–	–
C30	NV	nC(R)T + surgery	11.4 (8.2-15.5)	22.1 (12.7-23.7)**	–	–	9.3 (5.4-15.6)	14.7 (17.2-24.8)	–	–
C30	NV	Surgery	15.1 (11.9-19)	30.1 (8.2-15.5)***	14.9 (12.9-17.1)	16 (12.9-17.1)	8.4 (5.9-11.8)	16 (5.9-11.8)*	6.6 (3.5-12)	16.6 (3.5-12)*
C30	PA	nC(R)T + surgery	23.3 (17.8-29.8)	19.3 (17.8-29.8)	–	–	28.6 (21.3-37.1)	20.7 (17.8-28)	–	–
C30	PA	Surgery	23.9 (18.7-29.9)	31.6 (18.7-29.9)	–	–	22.1 (16.6-28.9)	22.9 (16.6-28.9)	10.7 (6.1-18.1)	18.9 (6.1-18.1)
C30	PFˆ	dCRT	80.3 (74.3-85.2)	62.7 (73-82.7)**	80 (74.4-84.6)	78.5 (74.4-84.6)	–	–	–	–
C30	PFˆ	nC(R)T + surgery	91 (86.3-94.2)	81.6 (74.3-85.2)*	–	–	92.6 (86.8-95.9)	80.4 (74.1-83.9)*	–	–
C30	PFˆ	Surgery	78.2 (73-82.7)	60.2 (86.3-94.2)***	79.6 (76.5-82.3)	75.8 (76.5-82.3)	82.5 (76.5-87.2)	77.5 (76.5-87.2)	92.5 (87.2-95.7)	81.4 (87.2-95.7)*
C30	RFˆ	dCRT	69.6 (61.7-76.5)	51 (61.6-74.1)**	69.9 (62.8-76.2)	78.3 (62.8-76.2)	–	–	–	–

Table 2 – *continued from previous page*

			Analysis 1		Analysis 2		Analysis 3		Analysis 4	
QLQ	**Sub-scale**	**Treatment**	**Baseline**	**Max. FU**	**Baseline**	**Max. FU**	**Baseline**	**Max. FU**	**Baseline**	**Max. FU**
C30	RF^	nC(R)T + surgery	80.4 (74-85.6)	57 (61.7-76.5)***	–	–	87 (79.8-91.9)	73.4 (54.1-66.1)*	–	–
C30	RF^	Surgery	68.2 (61.6-74.1)	38.5 (74-85.6)***	68.8 (64.6-72.7)	69.2 (64.6-72.7)	75.8 (69.5-81.1)	67.5 (69.5-81.1)	91.6 (85.9-95.2)	75 (85.9-95.2)**
C30	SF^	dCRT	75.6 (68.6-81.5)	64.2 (68-79.2)	75.6 (69.3-81.1)	81.9 (69.3-81.1)	–	–	–	–
C30	SF^	nC(R)T + surgery	76.9 (70-82.6)	69.2 (68.6-81.5)	–	–	82.4 (72.8-89.2)	80.2 (64-76.4)	–	–
C30	SF^	Surgery	74 (68-79.2)	52 (70-82.6)***	77.1 (73.4-80.5)	77.1 (73.4-80.5)	85.9 (80.8-89.7)	76.1 (80.8-89.7)*	82.8 (73.7-89.2)	79.4 (73.7-89.2)
OES	CH	nC(R)T + surgery	9.3 (3.9-20.5)	4.7 (13.9-41.5)	–	–	9.8 (2.1-35.8)	10.2 (2.6-14.8)	–	–
OES	CH	Surgery	25.2 (13.9-41.5)	21.5 (3.9-20.5)	–	–	24.3 (12.7-41.4)	9.2 (12.7-41.4)	–	–
OES	CG	dCRT	22.1 (14.4-32.5)	29.4 (12.7-23)	19.4 (12.8-28.2)	17.9 (12.8-28.2)	–	–	–	–
OES	CG	nC(R)T + surgery	15.5 (10.3-22.7)	24 (14.4-32.5)	–	–	17.4 (10.1-28.3)	19.3 (18.4-33.4)	–	–
OES	CG	Surgery	17.3 (12.7-23)	42.9 (10.3-22.7)***	16.8 (13.4-20.8)	23.9 (13.4-20.8)*	9.3 (6.1-13.7)	26.2 (6.1-13.7)***	–	–
OES	DM	dCRT	18.3 (12.9-25.5)	30.2 (15.3-23.4)*	15.5 (10.4-22.6)	27.1 (10.4-22.6)	–	–	–	–
OES	DM	nC(R)T + surgery	13.8 (10.1-18.7)	28 (12.9-25.5)**	–	–	12 (4.8-27)	20.8 (19.6-32.9)	–	–
OES	DM	Surgery	19 (15.3-23.4)	28.8 (10.1-18.7)*	17.8 (14.3-21.9)	21.8 (14.3-21.9)	12 (8.1-17.5)	22.2 (8.1-17.5)*	–	–
OES	DY	dCRT	57.9 (44-70.6)	47.5 (27.6-49.6)	–	–	–	–	–	–
OES	DY	nC(R)T + surgery	32.6 (22.9-44)	36.8 (44-70.6)	–	–	–	–	–	–
OES	DY	Surgery	38 (27.6-49.6)	33.5 (22.9-44)	–	–	–	–	–	–
OES	ED	nC(R)T + surgery	35.9 (28.7-43.8)	36.3 (25.4-34)	–	–	33.2 (24.7-43)	31.2 (29-41.4)	–	–
OES	ED	Surgery	29.5 (25.4-34)	42.6 (28.7-43.8)**	–	–	22.9 (18.4-28.2)	26.8 (18.4-28.2)	–	–
OES	PA	dCRT	30 (24.1-36.6)	19.9 (20.7-31.6)	25.5 (18.7-33.6)	7.2 (18.7-33.6)***	–	–	–	–
OES	PA	nC(R)T + surgery	25.7 (20.7-31.6)	26.6 (16.7-22.9)	–	–	31 (21.3-42.8)	14.4 (20.8-33.6)*	–	–
OES	PA	Surgery	19.6 (16.7-22.9)	18.4 (24.1-36.6)	20.8 (17.3-24.9)	15.9 (17.3-24.9)*	21.3 (16.1-27.7)	16.1 (16.1-27.7)	–	–
OES	RE	nC(R)T + surgery	12.5 (8.6-17.8)	11.7 (8.6-17.8)	–	–	13.2 (9-19)	21.7 (8.7-16.7)	–	–
OES	RE	Surgery	24.6 (18.5-31.9)	24.7 (18.5-31.9)	–	–	21.9 (15.4-30.2)	25.8 (15.4-30.2)	–	–
OES	SP	nC(R)T + surgery	4.4 (2.5-7.7)	7.5 (17.8-31.7)	–	–	2.9 (0.7-10.8)	10 (4.6-12.1)	–	–
OES	SP	Surgery	24 (17.8-31.7)	9.6 (2.5-7.7)**	–	–	32 (22.5-43.2)	7.6 (22.5-43.2)***	–	–
OES	SW	nC(R)T + surgery	14 (10.7-18.2)	13.4 (11.5-17.8)	–	–	13.1 (5.1-29.7)	12.8 (9.7-17.9)	–	–
OES	SW	Surgery	14.4 (11.5-17.8)	17 (10.7-18.2)	–	–	12 (8.2-17.2)	11.8 (8.2-17.2)	–	–
OES	TA	dCRT	35.9 (25.3-48)	44 (7.1-15.7)	35 (28.6-42)	21.2 (28.6-42)**	–	–	–	–
OES	TA	nC(R)T + surgery	10.4 (6.3-16.5)	44.1 (25.3-48)***	–	–	6.1 (3.2-11.1)	12 (29.1-45.3)	–	–
OES	TA	Surgery	10.7 (7.1-15.7)	38.1 (6.3-16.5)***	15.3 (13.3-17.4)	15.2 (13.3-17.4)	4.5 (2.8-7.1)	15.1 (2.8-7.1)**	–	–

Table 2 – *continued from previous page*

QLQ	Sub-scale	Treatment	Analysis 1		Analysis 2		Analysis 3		Analysis 4	
			Baseline	Max. FU	Baseline	Max. FU	Baseline	Max. FU	Baseline	Max. FU
STO22	BI	Surgery	–	–	–	–	–	–	7.7 (4.7-12.5)	29 (4.7-12.5)***
STO22	DM	Surgery	–	–	–	–	–	–	23.5 (16.1-32.9)	25 (16.1-32.9)
STO22	AL	Surgery	–	–	–	–	–	–	12.4 (4.7-29.2)	19.5 (4.7-29.2)
STO22	PA	Surgery	–	–	–	–	–	–	16.7 (10.5-25.4)	22.6 (10.5-25.4)
STO22	RE	Surgery	–	–	–	–	–	–	13.9 (7.4-24.4)	19.1 (7.4-24.4)
STO22	TA	Surgery	–	–	–	–	–	–	3.9 (2.1-7.2)	17 (2.1-7.2)**

Table 2: Overview of baseline- and maximum follow-up scores with 95% confidence interval in meta-analyses.
Analysis 1: short-term effect of neoadjuvant treatment, definitive treatment, and surgery alone on HRQoL for oesophageal cancer. Analysis 2: long-term effect of definitive treatment or surgery on HRQoL for oesophageal cancer. Analysis 3: long-term effect of neoadjuvant treatment with surgery or surgery alone on HRQoL for oesophageal cancer. Analysis 4: long-term effect of surgery on HRQoL for gastric cancer. The scores from all HRQoL subscales range from 0 to 100, with higher scores indicating worse HRQoL. In subscales annotated with '^' higher scores indicate better HRQoL. FU: follow-up; dCRT: definitive chemoradiotherapy; nC(R)T: neoadjuvant chemo(radio)therapy; QLQ: EORTC quality of life questionnaire; GHS: global health status; PF: physical functioning; RF: role functioning; EF: emotional functioning; CF: cognitive functioning; SF: social functioning; FA: fatigue; NV: nausea and vomiting; PA: pain; DY: dyspnoea; SL: insomnia; AP: appetite loss; CO: constipation; DI: diarrhoea; FI: financial difficulties; CH: choking while swallowing; CG: coughing; DM: dry mouth; ED: eating difficulties; RE: reflux; SP: speech; SW: trouble swallowing saliva; TA: taste; BI: body image; DM: dry mouth; AL: alopecia; *: $p<0.05$; **: $p<0.01$; ***: $p<0.001$.

Comparison between short-term HRQoL in oesophageal cancer patients receiving surgery vs. nC(R)T vs. dCRT

In the GHS analysis, respectively 10, 11, and 6 studies were included in the treatment groups surgery (N = 1,040), nC(R)T (N = 1,001), and dCRT (N = 482); see Figure 3A. Over the course of the first four months, a difference in GHS was observed between the treatment groups ($p < 0.001$), with nC(R)T showing a higher overall GHS score than dCRT (β = 0.55 (0.24-0.86), p = 0.002) and surgery (β = 0.42 (0.18-0.66), p = 0.002). Compared to baseline, GHS at the end of follow-up showed a clinically relevant decrease in the nC(R)T group (p = 0.008; 69.3 vs. 58.7), but not in the dCRT group (p = 0.137; 56.6 vs. 49.7) or surgery group (p = 0.055; 59.8 vs. 52.6); see Table 2. At the end of four months follow-up, no differences were identified between nC(R)T, dCRT, and surgery (p-values > 0.05).

In the subscale physical functioning, a difference between nC(R)T, dCRT, and surgery was also identified over the course of four months ($p < 0.001$), with nC(R)T showing better overall physical functioning compared to dCRT ($\beta = 0.91$ (0.34–1.48), $p = 0.004$) and surgery ($\beta = 1.03$ (0.51–1.56), $p < 0.001$); see Figure 4A.

In various HRQoL subscales, no differences were identified between nC(R)T, dCRT and surgery. When a difference was identified, nC(R)T demonstrated superiority to dCRT and surgery, except for financial worries in which surgery demonstrated better results compared to nC(R)T ($p = 0.019$); see Supplementary Figures 1.1–1.25 and Supplementary Table 3.

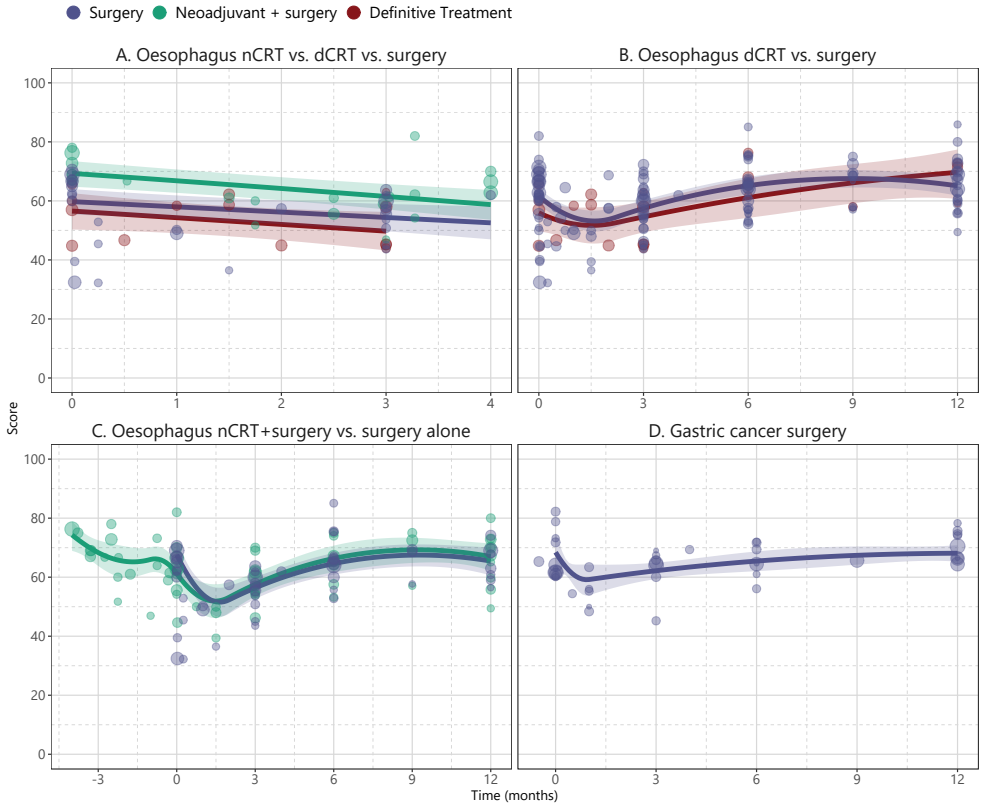


Figure 3: Results from the meta-analyses on patient-reported Global Health Status.
A: short-term effect of neoadjuvant treatment, definitive treatment, and surgery alone on HRQoL for oesophageal cancer. B: long-term effect of definitive treatment or surgery on HRQoL for oesophageal cancer. C: long-term effect of neoadjuvant treatment with surgery or surgery alone on HRQoL for oesophageal cancer. D: long-term effect of surgery on HRQoL for gastric cancer.

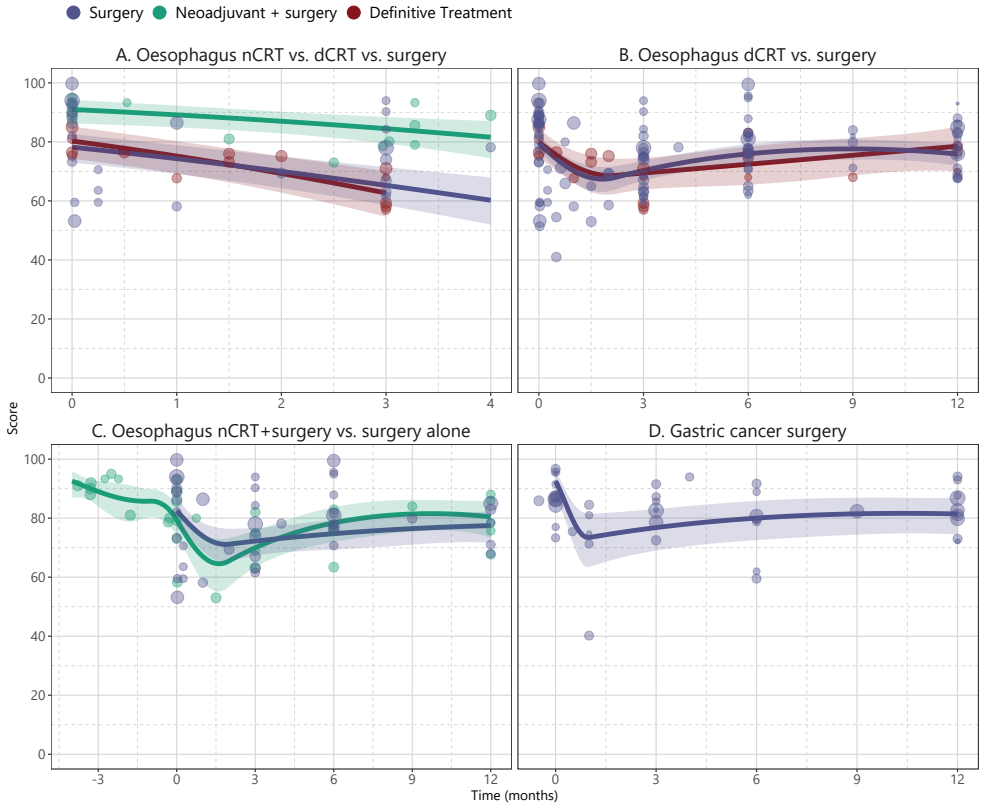


Figure 4: Results from the meta-analyses on patient-reported Physical Functioning.
A: short-term effect of neoadjuvant treatment, definitive treatment, and surgery alone on HRQoL for oesophageal cancer. B: long-term effect of definitive treatment or surgery on HRQoL for oesophageal cancer. C: long-term effect of neoadjuvant treatment with surgery or surgery alone on HRQoL for oesophageal cancer. D: long-term effect of surgery on HRQoL for gastric cancer.

Comparison of HRQoL in oesophageal cancer patients receiving surgery vs. dCRT

In the GHS analysis, 22 studies with patients who underwent surgery (N = 2,254), and six studies with patients who received dCRT (N = 482) were included (Figure 3B). Of the included studies on surgery, 62% included previously untreated patients, and 27% received pretreatment.

Over the course of 12 months, no difference in GHS was identified between surgery and dCRT ($p = 0.08$). GHS at 12 months demonstrated a clinically relevant increase compared to baseline in the dCRT group ($p = 0.003$; 56.0 vs. 69.7), but not in the surgery group ($p = 0.08$; 61.6 vs. 65.1); see Table 2. At 12 months, no difference in GHS score was identified between the dCRT or surgery group ($p = 0.33$).

In the subscale physical functioning, no difference was identified between the dCRT group and surgery group over the course of 12 months ($p = 0.87$); see Figure 4B. Physical functioning scores returned to baseline levels at 12 months for both the dCRT group ($p = 0.70$; 80.0 vs. 78.5) and the surgery group ($p = 0.19$, 79.6 vs. 75.8).

With respect to the other subscales, only differences between dCRT and surgery were identified in the subscales appetite ($p = 0.03$), and taste ($p < 0.001$), in which surgery demonstrated better scores compared to dCRT (Supplementary Figures 2.1–2.15).

Comparison of HRQoL in oesophageal cancer patients receiving nC(R)T with surgery vs. surgery alone

For the analysis of GHS, 11 studies were included with patients receiving nC(R)T and surgery (N = 1,015) compared to 10 studies with patients who underwent surgery alone (N = 1,021); see Figure 3C. Over the course of 12 months following surgery, no difference in GHS was identified between patients receiving nC(R)T with surgery or surgery alone ($p = 0.63$). No difference in GHS was identified at 12 months postoperatively, compared to onset of treatment, for both the nC(R)T with surgery group ($p = 0.13$; 74.3 vs. 67.0) and surgery alone group ($p = 0.58$; 67.6 vs. 65.4); see Table 2. Prior to surgery, the nC(R)T with surgery group experienced a clinically relevant decline of 13.1 points in GHS, and a further decline of 9.3 points until the lowest value reached. The surgery alone group experienced a clinically relevant decline in GHS of 16.0 points after surgery until the lowest value reached. Between both groups, no differences in GHS could be identified at 12 months ($p = 0.65$).

In the subscale physical functioning, no differences between both nC(R)T with surgery and surgery alone were identified over the course of 12 months ($p = 0.45$); see Figure 4C. At 12 months, physical functioning scores were not significantly different from baseline in the surgery alone group ($p = 0.24$; 82.5 vs. 77.5), whereas a significant decline was present in the nC(R)T with surgery group ($p = 0.04$; 92.6 vs. 80.4).
For the other subscales, both nC(R)T with surgery and surgery alone demonstrated similar results (p-values > 0.05). Only in the insomnia subscale patients receiving nC(R)T with surgery demonstrated better scores compared to surgery alone over the course of 12 months ($\beta = 0.50$ (0.08-0.92), $p = 0.02$). In the surgery alone group, the symptom scales fatigue ($p = 0.004$), dyspnoea ($p = 0.002$), diarrhoea ($p = 0.002$), dry mouth ($p = 0.05$), taste ($p = 0.001$) and coughing ($p < 0.001$) showed clinically relevant deterioration at 12 months compared to baseline scores, whereas speech ($p < 0.001$) improved significantly at 12 months compared to baseline. In the nC(R)T with surgery group, physical functioning ($p = 0.04$), role functioning ($p = 0.04$), and fatigue ($p = 0.015$) showed clinically relevant deterioration at 12 months compared to baseline, whereas the symptom scales appetite ($p = 0.004$), fatigue ($p = 0.015$), insomnia ($p = 0.05$), and pain ($p = 0.021$) showed clinically relevant improvement compared to baseline (Supplementary Figures 3.1–3.24).

HRQoL in gastric cancer patients receiving surgery
Nine studies describing GHS in patients with gastric cancer, treated with either surgery alone (N = 1,123) or surgery with adjuvant chemotherapy (N = 63) were included in these analyses (Figure 3D). Over the course of 12 months, the largest decline was seen at one month postoperatively (9.5 points), albeit not clinically relevant. At 12 months, GHS returned to baseline level ($p = 0.95$; 68.4 vs. 68.1).
Over the course of 12 months, the physical functioning subscale also showed the largest decline at one month postoperatively (19.1 points). At 12 months, physical functioning showed a significant and clinically relevant decrease compared to baseline ($p = 0.021$, 92.5 vs. 81.4); see Table 2 and Figure 4.
Role functioning ($p = 0.004$), and the symptom scales fatigue ($p = 0.03$), nausea/vomiting ($p = 0.04$), diarrhoea ($p < 0.001$), taste ($p = 0.001$), and body-image ($p < 0.001$) showed clinically relevant deterioration at 12 months compared to baseline scores (Supplementary Figures 4.1–4.20).

Geographic differences

We compared HRQoL outcomes from studies originating in Asian countries to studies from the Western World for oesophageal cancer (Figure 5). Since the majority of Asian studies only concerned surgery (N = 11, 65%), we could not compare geographical differences in HRQoL with chemoradiotherapy treatment. Likewise, the number of gastric cancer studies was not sufficient to compare HRQoL outcomes for gastric cancer. Overall, the Asian studies showed more impaired scores at baseline compared to Western studies. The Western studies showed an overall deterioration of the scores in the short-term following surgery and an increase thereafter, whereas the Asian studies demonstrated an improvement of scores directly following surgery. Statistically significant overall differences were observed in all EORTC subscales between the Asian and Western studies (Supplementary Figures 5.1–5.8).

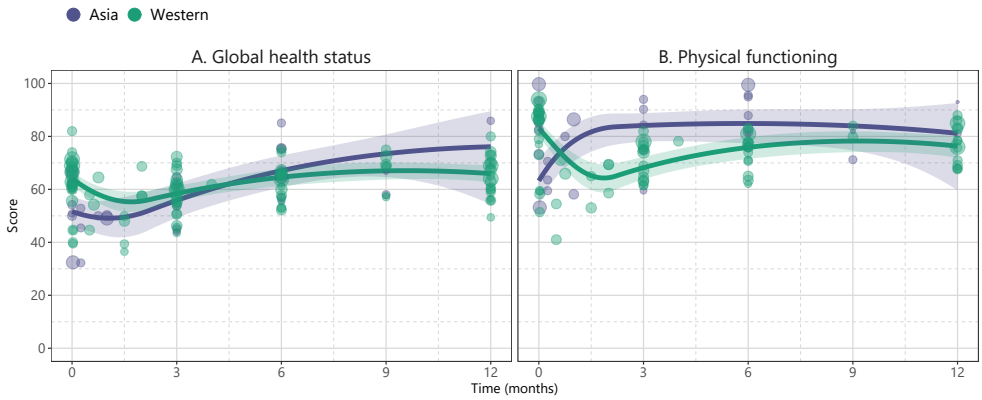


Figure 5: Results from the meta-analyses on HRQoL comparing studies conducted in Asia and in the Western world.
A: analysis comparing the effect from surgery on GHS for Asian vs. Western world studies over 12 months. B: analysis comparing the effect from surgery on physical functioning for Asian vs. Western world studies over 12 months.

Discussion

We systematically reviewed HRQoL data of curatively-treated oesophageal and gastric cancer patients, with the aim to compare the impact on HRQoL between different curative treatment strategies.

Over the course of the first four months after treatment initiation, patients who received nC(R)T prior to surgery demonstrated significantly better GHS compared to patients who received dCRT or surgery alone. It should, however, be taken into account that selection bias could play a role in this outcome, since younger and vital patients more often receive nC(R)T with surgery, compared to dCRT or surgery alone.[95] In the vast majority of HRQoL subscales, the same pattern was identified, with neoadjuvantly treated patients demonstrating significantly better results compared to patients receiving dCRT or surgery. Although other studies showed better HRQoL with dCRT compared to neoadjuvantly treated patients who also underwent a resection at short term[68], it should be noted that these studies included patients who received surgery following nC(R)T. In our analyses, we only focused on nC(R)T prior to surgery, since we wanted to demonstrate the effect of nC(R)T on short-term HRQoL, without the interference of surgery.

In order to analyse the influence of curative treatments on the long-term, we also examined HRQoL between patients who received nC(R)T and surgery compared to surgery alone over the first 12 months postoperatively. Overall, no differences were identified in HRQoL between these two treatments. Although a clinically-relevant impact of nC(R)T on HRQoL was observed prior to surgery, a further decline was seen upon resection in both the group receiving nC(R)T with surgery and the group receiving surgery alone. The HRQoL subscales showed similar scores for surgery alone and nC(R)T with surgery.

Likewise, no significant differences in HRQoL were identified between patients receiving dCRT or surgery over the first year following treatment initiation. HRQoL is often included in the decision-making regarding treatment when choosing for dCRT or surgery. As we did not find major differences in HRQoL in the first year following treatment initiation between both treatment options, the equivalence of both treatments in terms of HRQoL impact could be taken into account in clinical practice.

Whereas oesophagectomy caused a clinically relevant decline in GHS directly following surgery, the decline in GHS following gastrectomy was considered not clinically relevant. Overall, we found that both oesophagectomy and gastrectomy demonstrated long-lasting clinically relevant deterioration in various HRQoL functioning and symptom scales. For instance, the physical and role functioning remained impaired at 12 months following gastrectomy, and many symptom scales showed long-term deterioration, such as diarrhoea and taste for both gastrectomy and oesophagectomy. This has also been shown by other studies, which showed that HRQoL remained impaired on the long-term following surgery.[17,19,36,96] However, GHS scores in our meta-analyses showed no decline at one year following surgery in gastric cancer, as well as in oesophageal cancer.

We compared studies conducted in Asia with studies from the Western world, to identify potential structural differences between both geographical regions. Due to scarcity in HRQoL data available in both geographical regions, we only analysed the difference in HRQoL in patients with oesophageal cancer who received surgery. In nearly all subscales, a statistically significant difference was identified between Asian studies and studies form the Western world, with Asian studies demonstrating slightly better scores. As both the preferred treatment strategies and tumour characteristics vary between Asian countries and the Western world, a difference could be expected.[7,97,98] It is unclear whether these differences are related to culture, treatment, or other factors. Although the main meta-analyses included studies from all over the world, the differences identified between two regions suggest that these data may not be generalisable to the whole world, and should be interpreted with caution.

We performed an extensive blinded systematic search of both curatively treated oesophageal and gastric cancer, to our knowledge the first systematic review addressing both cancer types. Moreover, we included all studies reporting HRQoL, regardless of type of questionnaire applied, and we performed multiple robust meta-analyses. There are, however, also several limitations. First, we could not select studies based on quality, due to limited availability of high-quality HRQoL studies, a finding also demonstrated by others.[30,99] In this study, only 39% (N = 19) of the included studies were considered of high-quality. Many trials had no robust HRQoL reporting (N = 23), or showed selective reporting (N = 19). Several studies also introduced bias by subsequently excluding patients with serious complications, recurrence of disease, or patients who died during the first year after treatment, potentially

introducing optimism bias.[45,77,79,80] Included studies often did not report on questionnaire compliance[43,54,61–63,67,78,81], did not describe specific treatment provided[47,51,56,61,62,70,73], or did not report if patients were pre-treated prior to surgery[41,45,58,76]. Thus, current HRQoL studies are often of low-quality and individually have limited value in current clinical practice. By combining these individual studies in our meta-analyses, we aim to improve the clinical applicability of HRQoL, although the findings should be interpreted carefully. Secondly, another selection bias was potentially introduced, as the inclusion periods of the studies spanned a large range of time. Although we excluded studies published prior to 2000 as they examined outdated surgical techniques, studies performed in the early 21st century also differ substantially in surgical techniques compared to the studies published later on.[100,101] As further selection on inclusion period was not possible due to data scarcity, this could have potentially influenced the outcome of the impact of surgery on HRQoL.

Whereas many studies have been published on HRQoL regarding surgery, there is a scarcity of data regarding HRQoL in patients treated with multimodality treatment, while this is most frequently used in current clinical practice.[3] Especially in patients with curatively-treated gastric cancer, a lack of HRQoL data was identified, and we were thus unable to determine the impact of perioperative or adjuvant chemoradiotherapy on HRQoL in these patients. Thus, there is a need for more studies that focus on HRQoL. As randomisation is often not ethical, and many non-randomised studies do not account for prognostic imbalance within different groups, these data have limitations when applied in clinical practice. Thus, to generate more reliable data, we recommend future randomised controlled studies to also focus on HRQoL data instead on treatment comparisons alone. In addition, we recommend these studies to use the CONSORT PRO guidelines for robust reporting and interpretation of HRQoL data.[102,103]

To conclude, although short-term differences were identified between curative treatment options, no differences were identified in HRQoL between surgery alone and nC(R)T with surgery, or between surgery alone and dCRT over the first year in oesophageal cancer patients. In gastric cancer patients, surgery did not show a lasting clinically relevant deterioration of HRQoL. Furthermore, there is a need for more prospective, high-quality studies on HRQoL in oesophageal and gastric cancer, reflecting contemporary clinical practice.

References

1. Bray, F., Ferlay, J., Soerjomataram, I., *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians 68, 394–424 (2018).
2. Hulscher, J. B., van Sandick, J. W., de Boer, A. G., *et al.* Extended Transthoracic Resection Compared with Limited Transhiatal Resection for Adenocarcinoma of the Esophagus. New England Journal of Medicine 347, 1662–1669 (2002).
3. Van Hagen, P., Hulshof, M., van Lanschot, J., *et al.* Preoperative Chemoradiotherapy for Esophageal or Junctional Cancer. New England Journal of Medicine 366, 2074–2084 (2012).
4. Kelsen, D. P., Ginsberg, R., Pajak, T. F., *et al.* Chemotherapy Followed by Surgery Compared with Surgery Alone for Localized Esophageal Cancer. New England Journal of Medicine 339, 1979–1984 (1998).
5. Badgwell, B., Das, P. & Ajani, J. Treatment of localized gastric and gastroesophageal adenocarcinoma: The role of accurate staging and preoperative therapy. Journal of Hematology and Oncology 10, 1–7 (2017).
6. Omloo, J. M., Lagarde, S. M., Hulscher, J. B., *et al.* Extended transthoracic resection compared with limited transhiatal resection for adenocarcinoma of the mid/distal esophagus: Five-year survival of a randomized clinical trial. Annals of Surgery 246, 992–1000 (2007).
7. Zhang, H. Z., Jin, G. F. & Shen, H. B. Epidemiologic differences in esophageal cancer between Asian and Western populations. Chinese journal of cancer 31, 281–286 (2012).
8. Al-Batran, S. E., Homann, N., Pauligk, C., *et al.* Perioperative chemotherapy with fluorouracil plus leucovorin, oxaliplatin, and docetaxel versus fluorouracil or capecitabine plus cisplatin and epirubicin for locally advanced, resectable gastric or gastro-oesophageal junction adenocarcinoma (FLOT4): a randomised, phase 2/3 trial. The Lancet 393, 1948–1957 (2019).
9. Cunningham, D., Allum, W. H., Stenning, S. P., *et al.* Perioperative Chemotherapy versus Surgery Alone for Resectable Gastroesophageal Cancer. New England Journal of Medicine 355, 11–20 (2006).
10. Sasako, M., Sakuramoto, S., Katai, H., *et al.* Five-year outcomes of a randomized phase III trial comparing adjuvant chemotherapy with S-1 versus surgery alone in stage II or III gastric cancer. Journal of Clinical Oncology 29, 4387–4393 (2011).
11. Smalley, S. R., Benedetti, J. K., Haller, D. G., *et al.* Updated analysis of SWOG-directed intergroup study 0116: A phase III trial of adjuvant radiochemotherapy versus observation after curative gastric cancer resection. Journal of Clinical Oncology 30, 2327–2333 (2012).
12. Tepper, J., Krasna, M. J., Niedzwiecki, D., *et al.* Phase III trial of trimodality therapy with cisplatin, fluorouracil, radiotherapy, and surgery compared with surgery alone for esophageal cancer: CALGB 9781. Journal of Clinical Oncology 26, 1086–1092 (2008).
13. Ando, N., Kato, H., Igaki, H., *et al.* A randomized trial comparing postoperative adjuvant chemotherapy with cisplatin and 5-fluorouracil versus preoperative chemotherapy for localized advanced squamous cell carcinoma of the thoracic esophagus (JCOG9907). Annals of Surgical Oncology 19, 68–74 (2012).
14. Kitagawa, Y., Uno, T., Oyama, T., *et al.* Esophageal cancer practice guidelines 2017 edited by the Japan Esophageal Society: part 1. Esophagus 16 (2019).

15. Scarpa, M., Valente, S., Alfieri, R., *et al.* Systematic review of health-related quality of life after esophagectomy for esophageal cancer. World Journal of Gastroenterology 17, 4660–4674 (2011).
16. Conroy, T., Marchal, F. & Blazeby, J. M. Quality of life in patients with oesophageal and gastric cancer: An overview. Oncology 70, 391–402 (2007).
17. Noordman, B. J., Verdam, M. G., Lagarde, S. M., *et al.* Effect of neoadjuvant chemoradiotherapy on health-Related quality of life in esophageal or junctional cancer: Results from the randomized CROSS trial. Journal of Clinical Oncology 36, 268–275 (2018).
18. Safieddine, N., Xu, W., Quadri, S. M., *et al.* Health-related quality of life in esophageal cancer: Effect of neoadjuvant chemoradiotherapy followed by surgical intervention. Journal of Thoracic and Cardiovascular Surgery 137, 36–42 (2009).
19. Derogar, M., Orsini, N., Sadr-Azodi, O. & Lagergren, P. Influence of major postoperative complications on health-related quality of life among long-term survivors of esophageal cancer surgery. Journal of Clinical Oncology 30, 1615–1619 (2012).
20. Jacobs, M., Macefield, R. C., Elbers, R. G., *et al.* Meta-analysis shows clinically relevant and long-lasting deterioration in health-related quality of life after esophageal cancer surgery. Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation 23, 1155–1176 (2014).
21. Djärv, T., Metcalfe, C., Avery, K. N., *et al.* Prognostic value of changes in health-related quality of life scores during curative treatment for esophagogastric cancer. Journal of Clinical Oncology 28, 1666–1670 (2010).
22. Djärv, T. & Lagergren, P. Six-month postoperative quality of life predicts long-term survival after oesophageal cancer surgery. European Journal of Cancer 47, 530–535 (2011).
23. Alghamedi, A., Buduhan, G., Tan, L., *et al.* Quality of life assessment in esophagectomy patients. Annals of Translational Medicine 6, 84–84 (2018).
24. Quinten, C., Coens, C., Mauer, M., *et al.* Baseline quality of life as a prognostic indicator of survival: a meta-analysis of individual patient data from EORTC clinical trials. The Lancet Oncology 10, 865–871 (2009).
25. Liberati, A., Altman, D. G., Tetzlaff, J., *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Journal of clinical epidemiology 62, e1–e34 (2009).
26. Siewert, J. R. & Stein, H. J. Classification of adenocarcinoma of the oesophagogastric junction. British Journal of Surgery 85, 1457–1459 (1998).
27. Higgins, J. P. & Green, S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.3.0 (Chichester, 2015).
28. Efficace, F., Bottomley, A., Osoba, D., *et al.* Beyond the development of health-related quality-of-life (HRQOL) measures: A checklist for evaluating HRQOL outcomes in cancer clinical trials - Does HRQOL evaluation in prostate cancer research inform clinical decision making? Journal of Clinical Oncology 21, 3502–3511 (2003).
29. Stang, A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. European Journal of Epidemiology 25, 603–605 (2010).
30. Jacobs, M., Macefield, R. C., Blazeby, J. M., *et al.* Systematic review reveals limitations of studies evaluating health-related quality of life after potentially curative treatment for esophageal cancer. Quality of Life Research 22, 1787–1803 (2013).
31. Noordman, B. J., Wijnhoven, B. P., Lagarde, S. M., *et al.* Neoadjuvant chemoradiotherapy plus surgery versus active surveillance for oesophageal cancer: A stepped-wedge cluster randomised trial. BMC Cancer 18, 1–12 (2018).

32. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2018.
33. Viechtbauer, W. Conducting meta-analyses in R with the metafor. Journal of Statistical Software 36, 1–48 (2010).
34. Bozdogan, H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika 52, 345–370 (1987).
35. Holm, S. A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics, 65–70 (1979).
36. Noordman, B. J., Verdam, M. G., Lagarde, S. M., *et al.* Impact of neoadjuvant chemoradiotherapy on health-related quality of life in long-term survivors of esophageal or junctional cancer: Results from the randomized CROSS trial. Annals of Oncology 29, 445–451 (2018).
37. Cavallin, F., Pinto, E., Saadeh, L. M., *et al.* Health related quality of life after oesophagectomy: Elderly patients refer similar eating and swallowing difficulties than younger patients. BMC Cancer 15, 1–9 (2015).
38. Cense, H. A., Hulscher, J. B., De Boer, A. G., *et al.* Effects of prolonged intensive care unit stay on quality of life and long-term survival after transthoracic esophageal resection. Critical Care Medicine 34, 354–362 (2006).
39. Chou, S. H., Li, H. P., Lee, J. Y., *et al.* Radical resection or chemoradiotherapy for cervical esophageal cancer? World Journal of Surgery 34, 1832–1839 (2010).
40. De Boer, A. G., Van Lanschot, J. J., Van Sandick, J. W., *et al.* Quality of life after transhiatal compared with extended transthoracic resection for adenocarcinoma of the esophagus. Journal of Clinical Oncology 22, 4202–4208 (2004).
41. Egberts, J. H., Schniewind, B., Bestmann, B., *et al.* Impact of the site of anastomosis after oncologic esophagectomy on quality of life - A prospective, longitudinal outcome study. Annals of Surgical Oncology 15, 566–575 (2008).
42. Fang, F. M., Tsai, W. L., Chiu, H. C., *et al.* Quality of life as a survival predictor for esophageal squamous cell carcinoma treated with radiotherapy. International Journal of Radiation Oncology Biology Physics 58, 1394–1404 (2004).
43. Garant, A., Whitaker, T. J., Spears, G. M., *et al.* A Comparison of Patient-Reported Health-Related Quality of Life During Proton Versus Photon Chemoradiation Therapy for Esophageal Cancer. Practical Radiation Oncology 9, 410–417 (2019).
44. Gillham, C. M., Aherne, N., Rowley, S., *et al.* Quality of Life and Survival in Patients Treated with Radical Chemoradiation Alone for Oesophageal Cancer. Clinical Oncology 20, 227–233 (2008).
45. Gradauskas, P., Rubikas, R. & Saferis, V. Changes in quality of life after esophageal resections for carcinoma. Medicina (Kaunas, Lithuania) 42, 187–194 (2006).
46. Haj Mohammad, N., De Rooij, S., Hulshof, M., *et al.* Activities of daily living and quality of life during treatment with neoadjuvant chemoradiotherapy and after surgery in patients with esophageal cancer. Journal of Surgical Oncology 114, 684–690 (2016).
47. Hauser, C., Patett, C., von Schoenfels, W., *et al.* Does neoadjuvant treatment before oncologic esophagectomy affect the postoperative quality of life? A prospective, longitudinal outcome study. Diseases of the Esophagus 28, 652–659 (2015).
48. Huang, Q., Zhong, J., Yang, T., *et al.* Impacts of anastomotic complications on the health-related quality of life after esophagectomy. Journal of Surgical Oncology 111, 365–370 (2015).
49. Kachnic, L. A., Winter, K., Wasserman, T., *et al.* Longitudinal quality-of-life analysis of RTOG 94-05 (Int 0123): A phase III trial of definitive chemoradiotherapy for esophageal cancer. Gastrointestinal Cancer Research 4, 45–52 (2011).

50. Lv, Y., Zhang, J. & Qiao, L. Quality of life in patients with esophageal cancer receiving definitive chemoradiotherapy or esophagectomy. Molecular and Clinical Oncology 2, 870–874 (2014).
51. Malmström, M., Klefsgard, R., Ivarsson, B., *et al.* Quality of life measurements as an indicator for timing of support after oesophagectomy for cancer: A prospective study. BMC Health Services Research 15, 1–7 (2015).
52. Nafteux, P., Durnez, J., Moons, J., *et al.* Assessing the relationships between health-related quality of life and postoperative length of hospital stay after oesophagectomy for cancer of the oesophagus and the gastro-oesophageal junction. European Journal of Cardio-thoracic Surgery 44, 525–533 (2013).
53. Parameswaran, R., Blazeby, J. M., Hughes, R., *et al.* Health-related quality of life after minimally invasive oesophagectomy. British Journal of Surgery 97, 525–531 (2010).
54. Ramakrishnaiah, V. P., Dash, N. R., Pal, S., *et al.* Quality of life after oesophagectomy in patients with carcinoma of oesophagus: A prospective study. Indian Journal of Cancer 51, 346–351 (2014).
55. Rees, J., Hurt, C. N., Gollins, S., *et al.* Patient-reported outcomes during and after definitive chemoradiotherapy for oesophageal cancer. British Journal of Cancer 113, 603–610 (2015).
56. Reynolds, J. V., McLaughlin, R., Moore, J., *et al.* Prospective evaluation of quality of life in patients with localized oesophageal cancer treated by multimodality therapy or surgery alone. British Journal of Surgery 93, 1084–1090 (2006).
57. Alderson, D., Cunningham, D., Nankivell, M., *et al.* Neoadjuvant cisplatin and fluorouracil versus epirubicin, cisplatin, and capecitabine followed by resection in patients with oesophageal adenocarcinoma (UK MRC OE05): an open-label, randomised phase 3 trial. The Lancet Oncology 18, 1249–1260 (2017).
58. Sarkaria, I. S., Rizk, N. P., Goldman, D. A., *et al.* Early Quality of Life Outcomes After Robotic-Assisted Minimally Invasive and Open Esophagectomy. Annals of Thoracic Surgery 108, 920–928 (2019).
59. Sugawara, K., Yoshimura, S., Yagi, K., *et al.* Long-term health-related quality of life following robot-assisted radical transmediastinal esophagectomy. Surgical Endoscopy 34, 1602–1611 (2019).
60. Sunde, B., Klevebro, F., Johar, A., *et al.* Health-related quality of life in a randomized trial of neoadjuvant chemotherapy or chemoradiotherapy plus surgery in patients with oesophageal cancer (NeoRes trial). British Journal of Surgery 106, 1452–1463 (2019).
61. Tatematsu, N., Ezoe, Y., Tanaka, E., *et al.* Impact of neoadjuvant chemotherapy on physical fitness, physical activity, and health-related quality of life of patients with resectable esophageal cancer. American Journal of Clinical Oncology: Cancer Clinical Trials 36, 53–56 (2013).
62. Tatematsu, N., Hasegawa, S., Tanaka, E., *et al.* Impact of oesophagectomy on physical fitness and health-related quality of life in patients with oesophageal cancer. European Journal of Cancer Care 22, 308–313 (2013).
63. Van der Sluis, P. C., van der Horst, S., May, A. M., *et al.* Robot-assisted Minimally Invasive Thoracolaparoscopic Esophagectomy Versus Open Transthoracic Esophagectomy for Resectable Esophageal Cancer: A Randomized Controlled Trial. Annals of surgery 269, 621–630 (2019).
64. Van Heijl, M., Sprangers, M. A., De Boer, A. G., *et al.* Preoperative and early postoperative quality of life predict survival in potentially curable patients with esophageal cancer. Annals of Surgical Oncology 17, 23–30 (2010).

65. Van Meerten, E., van der Gaast, A., Looman, C. W., *et al.* Quality of Life During Neoadjuvant Treatment and After Surgery for Resectable Esophageal Carcinoma. International Journal of Radiation Oncology Biology Physics 71, 160–166 (2008).
66. Yamashita, H., Omori, M., Okuma, K., *et al.* Longitudinal assessments of quality of life and late toxicities before and after definitive chemoradiation for esophageal cancer. Japanese Journal of Clinical Oncology 44, 78–84 (2014).
67. Zeng, J. & Liu, J. S. Quality of life after three kinds of esophagectomy for cancer. World Journal of Gastroenterology 18, 5106–5113 (2012).
68. Avery, K. N., Metcalfe, C., Barham, C. P., *et al.* Quality of life during potentially curative treatment for locally advanced oesophageal cancer. British Journal of Surgery 94, 1369–1376 (2007).
69. Barbour, A. P., Lagergren, P., Hughes, R., *et al.* Health-related quality of life among patients with adenocarcinoma of the gastro-oesophageal junction treated by gastrectomy or oesophagectomy. British Journal of Surgery 95, 80–84 (2008).
70. Blazeby, J. M., Conroy, T., Hammerlid, E., *et al.* Clinical and psychometric validation of an EORTC questionnaire module, the EORTC QLQ-OES18, to assess quality of life in patients with oesophageal cancer. European Journal of Cancer 39, 1384–1394 (2003).
71. Blazeby, J. M., Sanford, E., Falk, S. J., *et al.* Health-related quality of life during neoadjuvant treatment and surgery for localized esophageal carcinoma. Cancer 103, 1791–1799 (2005).
72. Bonnetain, F., Bouché, O., Michel, P., *et al.* A comparative longitudinal quality of life study using the Spitzer quality of life index in a randomized multicenter phase III trial (FFCD 9102): Chemoradiation followed by surgery compared with chemoradiation alone in locally advanced squamous resectable. Annals of Oncology 17, 827–834 (2006).
73. Brooks, J. A., Kesler, K. A., Johnson, C. S., *et al.* Prospective analysis of quality of life after surgical resection for esophageal cancer: Preliminary results. Journal of Surgical Oncology 81, 185–194 (2002).
74. Blazeby, J. M., Conroy, T., Bottomley, A., *et al.* Clinical and psychometric validation of a questionnaire module, the EORTC QLQ-STO 22, to assess quality of life in patients with gastric cancer. European Journal of Cancer 40, 2260–2268 (2004).
75. Kassam, Z., Mackay, H., Buckley, C. A., *et al.* Evaluating the impact on quality of life of chemoradiation in gastric cancer. Current Oncology 17, 77–84 (2010).
76. Kim, A. R., Cho, J., Hsu, Y. J., *et al.* Changes of quality of life in gastric cancer patients after curative resection: A longitudinal cohort study in Korea. Annals of Surgery 256, 1008–1013 (2012).
77. Kobayashi, D., Kodera, Y., Fujiwara, M., *et al.* Assessment of quality of life after gastrectomy using EORTC QLQ-C30 and STO22. World Journal of Surgery 35, 357–364 (2011).
78. Kong, H., Kwon, O. K. & Yu, W. Changes of quality of life after gastric cancer surgery. Journal of Gastric Cancer 12, 194–200 (2012).
79. Li, J., Liu, H. & Yang, G. Quality of life after esophagogastrostomy plus gastrojejunostomy reconstruction following proximal gastrectomy: A comparative study of three surgical procedures. International Journal of Clinical and Experimental Medicine 11, 9779–9790 (2018).
80. Park, K. B., Park, J. Y., Lee, S. S., *et al.* Impact of body mass index on the quality of life after total gastrectomy for gastric cancer. Cancer Research and Treatment 50, 852–860 (2018).
81. Pruthi, D., Ahmad, M., Gupta, M., *et al.* Assessment of quality of life in resectable gastric cancer patients undergoing chemoradiotherapy as adjuvant treatment. South Asian Journal of Cancer 7, 16–20 (2018).

82. Wu, C. W., Chiou, J. M., Ko, F. S., *et al.* Quality of life after curative gastrectomy for gastric cancer in a randomised controlled trial. British Journal of Cancer 98, 54–59 (2008).
83. Aaronson, N. K., Ahmedzai, S., Bergman, B., *et al.* The European organization for research and treatment of cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. Journal of the National Cancer Institute 85, 365–376 (1993).
84. Blazeby, J. M., Alderson, D., Winstone, K., *et al.* Development of an EORTC questionnaire module to be used in quality of life assessment for patients with oesophageal cancer. European Journal of Cancer Part A 32, 1912–1917 (1996).
85. Darling, G., Eton, D. T., Sulman, J., *et al.* Validation of the functional assessment of cancer therapy esophageal cancer subscale. Cancer 107, 854–863 (2006).
86. List, M. A., D'Antonio, L. L., Cella, D. F., *et al.* The performance status scale for head and neck cancer patients and the functional assessment of cancer therapy-head and neck scale: A study of utility and validity. Cancer 77, 2294–2301 (1996).
87. Stewart, A. L., Hays, R. D. & Ware, J. E. The MOS short-form general health survey: Reliability and validity in a patient population. Medical Care 26, 724–735 (1988).
88. De Haes, J. C. & Olschewski, M. Quality of life assessment in a cross-cultural context: Use of the Rotterdam Symptom Checklist in a multinational randomised trial comparing CMF and Zoladex (Goserlin) treatment in early breast cancer. Annals of Oncology 9, 745–750 (1998).
89. Gold, J. Quality of life measurements in patients with malignant disease. Journal of the Royal Society of Medicine 79, 622 (1986).
90. McNair, D. M., Lorr, M. & Droppleman, L. F. Manual profile of mood states (1971).
91. Noerholm, V., Groenvold, M., Watt, T., *et al.* Quality of life in the Danish general population - Normative data and validity of WHOQOL-BREF using Rasch and item response theory models. Quality of Life Research 13, 531–540 (2004).
92. Biermann, C. W., Schmidt, C. & Küchler, T. Development of a life quality questionnaire in bladder cancer surgery. British Journal of Urology 76, 412–413 (1995).
93. Korenaga, D., Orita, H., Okuyama, T., *et al.* Quality of life after gastrectomy in patients with carcinoma of the stomach. British Journal of Surgery 79, 248–250 (1992).
94. Barbour, A. P., Cormack, O. M., Baker, P. J., *et al.* Long-term health-related quality of life following esophagectomy: A nonrandomized comparison of thoracoscopically assisted and open surgery. Annals of Surgery 265, 1158–1165 (2017).
95. Sun, L., Zhao, F., Zeng, Y. & Yi, C. Risks and benefits of multimodal esophageal cancer treatments: A meta-analysis. Medical Science Monitor 23, 889–910 (2017).
96. Schandl, A., Lagergren, J., Johar, A. & Lagergren, P. Health-related quality of life 10 years after oesophageal cancer surgery. European Journal of Cancer 69, 43–50 (2016).
97. Van Rijswijk, A. S., Hagens, E. R., van der Peet, D. L., *et al.* Differences in Esophageal Cancer Surgery in Terms of Surgical Approach and Extent of Lymphadenectomy: Findings of an International Survey. Annals of Surgical Oncology, 2063–2072 (2019).
98. Karimi, P., Islami, F., Anandasabapathy, S., *et al.* Gastric cancer: Descriptive epidemiology, risk factors, screening, and prevention. Cancer Epidemiology Biomarkers and Prevention 23, 700–713 (2014).
99. Van Kleef, J. J., Ter Veer, E., Van den Boorn, H. G., *et al.* Quality of Life during Palliative Systemic Therapy for Esophagogastric Cancer: Systematic Review and Meta-Analysis. Journal of the National Cancer Institute 112, 12–29 (2019).
100. D'Amico, T. A. Outcomes after surgery for esophageal cancer. Gastrointestinal cancer research : GCR 1, 188–96 (2007).

101. Hofstetter, W., Swisher, S. G., Correa, A. M., *et al.* Treatment outcomes of resected esophageal cancer. Annals of Surgery 236, 376–385 (2002).
102. Calvert, M., Blazeby, J., Altman, D. G., *et al.* Reporting of patient-reported outcomes in randomized trials: The CONSORT PRO extension. JAMA - Journal of the American Medical Association 309, 814–822 (2013).
103. Calvert, M., Kyte, D., Mercieca-Bebber, R., *et al.* Guidelines for inclusion of patient-reported outcomes in clinical trial protocols the spirit-pro extension. JAMA - Journal of the American Medical Association 319, 483–494 (2018).

Supplement

Supplementary Figures are available on the website of Critical Reviews in Oncology / Hematology via: https://doi.org/10.1016/j.critrevonc.2020.103069.

Database	Search terms
PUBMED/	"Stomach Neoplasms"[Mesh] OR gastric cancer[tiab] OR gastric tumo*[tiab] OR gastric neoplasm* [tiab] OR gastric malig*[tiab] OR stomach cancer[tiab] OR stomach tumo*[tiab] OR stomach neoplasm*[tiab] OR stomach malig*[tiab] "Gastrectomy"[Mesh] OR "Combined Modality Therapy"[Mesh] OR surger*[tiab] OR surgical[tiab] OR resection*[tiab] OR gastrectom*[tiab] OR combined therap*[tiab] OR combined modality therap*[tiab] OR neoadjuvant[tiab] OR peri-operative[tiab] OR adjuvant[tiab] OR radiotherap*[tiab] OR radiation*[tiab] OR chemotherap*[tiab] OR systemic[tiab] OR chemoradiat*[tiab] "Esophageal Neoplasms"[Mesh] OR esophageal cancer[tiab] OR esophageal tumo*[tiab] OR esophageal neoplasm* [tiab] OR esophageal malig*[tiab] OR oesophageal cancer[tiab] OR oesophageal tumo*[tiab] OR oesophageal neoplasm* [tiab] OR oesophageal malig*[tiab] "Esophagectomy"[Mesh] OR "Combined Modality Therapy"[Mesh] OR surger*[tiab] OR surgical[tiab] OR resection*[tiab] OR esophagectom*[tiab] OR oesophagectom*[tiab] OR combined therap*[tiab] OR combined modality therap*[tiab] OR neoadjuvant[tiab] OR peri-operative[tiab] OR adjuvant[tiab] OR radiotherap*[tiab] OR radiation*[tiab] OR chemotherap*[tiab] OR systemic[tiab] OR chemoradiat*[tiab] "Quality of Life"[Mesh] OR "Patient Reported Outcome Measures"[Mesh] OR quality of life[tiab] OR qol [tiab] OR hrqol [tiab] OR hrql [tiab] OR QLQ-C30[tiab] OR EQ-5D[tiab] OR patient reported outcome*[tiab] OR questionnaire*[tiab] OR PROMS[tiab]
EMBASE	exp stomach tumor/ or ((gastric* or stomach*) adj3 (cancer* or tumo* or neoplasm* or malig*)).ti,ab,kw. exp gastrectomy/ or multimodality cancer therapy/ or (surger* or surgical or resection* or gastrectom* or combined therap* or combined modality therap* or neoadjuvant or peri-operative or adjuvant or radiotherap* or radiation* or chemotherap* or systemic or chemoradiat*).ti,ab,kw. exp esophagus tumor/ or ((esophag* or oesophag*) adj3 (cancer* or tumo* or neoplasm* or malig*)).ti,ab,kw. esophagus resection/ or esophagus surgery/ or multimodality cancer therapy/ or (surger* or surgical or resection* or esophagectom* or oesophagectom* or combined therap* or combined modality therap* or neoadjuvant or peri-operative or adjuvant or radiotherap* or radiation* or chemotherap* or systemic or chemoradiat*).ti,ab,kw. exp "quality of life"/ or patient-reported outcome/ or (quality of life or qol or hrqol or hrql or QLQ-C30 or EQ-5D or patient reported outcome* or questionnaire* or PROMS).ti,ab,kw.
PSYCINFO	(exp Stomach/ and exp Neoplasms/) or ((gastric* or stomach*) and (cancer* or tumo* or neoplasm* or malig*)).ti,ab,id. exp surgery/ or (surger* or surgical or resection* or gastrectom* or combined therap* or combined modality therap* or neoadjuvant or peri-operative or adjuvant or radiotherap* or radiation* or chemotherap* or systemic or chemoradiat*).ti,ab,id. (esophagus/ and exp neoplasms/) or ((esophag* or oesophag*) and (cancer* or tumo* or neoplasm* or malig*)).ti,ab,id. exp surgery/ or (surger* or surgical or resection* or esophagectom* or oesophagectom* or combined therap* or combined modality therap* or neoadjuvant or peri-operative or adjuvant or radiotherap* or radiation* or chemotherap* or systemic or chemoradiat*).ti,ab,id. "quality of life"/ or exp questionnaires/ or (quality of life or qol or hrqol or hrql or QLQ-C30 or EQ-5D or patient reported outcome* or questionnaire* or PROMS).ti,ab,id.

Supplementary Table 1 – *continued from previous page*

Database	Search terms
CINAHL	(MH "Stomach Neoplasms") OR (TI ((gastric* or stomach*) N3 (cancer* or tumo* or neoplasm* or malig*)) OR AB ((gastric* or stomach*) N3 (cancer* or tumo* or neoplasm* or malig*))) (MH "Gastrectomy") OR ((MH "Combined Modality Therapy+") OR TI (surger* or surgical or resection* or gastrectom* or combined therap* or combined modality therap* or neoadjuvant or peri-operative or adjuvant or radiotherap* or radiation* or chemotherap* or systemic or chemoradiat*) OR AB (surger* or surgical or resection* or gastrectom* or combined therap* or combined modality therap* or neoadjuvant or peri-operative or adjuvant or radiotherap* or radiation* or chemotherap* or systemic or chemoradiat*)) (MH "Esophageal Neoplasms") OR TI ((esophag* or oesophag*) N3 (cancer* or tumo* or neoplasm* or malig*)) OR AB ((esophag* or oesophag*) N3 (cancer* or tumo* or neoplasm* or malig*)) (MH "Combined Modality Therapy+") OR TI (surger* or surgical or resection* or esophagectom* or oesophagectom* or combined therap* or combined modality therap* or neoadjuvant or peri-operative or adjuvant or radiotherap* or radiation* or chemotherap* or systemic or chemoradiat*) OR AB (surger* or surgical or resection* or esophagectom* or oesophagectom* or combined therap* or combined modality therap* or neoadjuvant or peri-operative or adjuvant or radiotherap* or radiation* or chemotherap* or systemic or chemoradiat*) (MH "Quality of Life+") OR (MH "Patient-Reported Outcomes") OR TI (quality of life or qol or hrqol or hrql or QLQ-C30 or EQ-5D or patient reported outcome* or questionnaire* or PROMS) OR AB (quality of life or qol or hrqol or hrql or QLQ-C30 or EQ-5D or patient reported outcome* or questionnaire* or PROMS)
Cochrane Central registrer of Controlled Trials	MeSH descriptor: [Stomach Neoplasms] explode all trees ((gastric* or stomach*) near/3 (cancer* or tumo* or neoplasm* or malig*)):ti,ab,kw (Word variations have been searched) MeSH descriptor: [Gastrectomy] explode all trees MeSH descriptor: [Combined Modality Therapy] explode all trees (surger* or surgical or resection* or gastrectom* or combined therap* or combined modality therap* or neoadjuvant or peri-operative or adjuvant or radiotherap* or radiation* or chemotherap* or systemic or chemoradiat*):ti,ab,kw (Word variations have been searched) MeSH descriptor: [Esophageal Neoplasms] explode all trees ((esophag* or oesophag*) NEAR/3 (cancer* or tumo* or neoplasm* or malig*)):ti,ab,kw (Word variations have been searched) MeSH descriptor: [Esophagectomy] explode all trees MeSH descriptor: [Combined Modality Therapy] explode all trees (surger* or surgical or resection* or esophagectom* or oesophagectom* or combined therap* or combined modality therap* or neoadjuvant or peri-operative or adjuvant or radiotherap* or radiation* or chemotherap* or systemic or chemoradiat*):ti,ab,kw (Word variations have been searched) MeSH descriptor: [Quality of Life] explode all trees MeSH descriptor: [Patient Reported Outcome Measures] explode all trees (quality of life or qol or hrqol or hrql or QLQ-C30 or EQ-5D or patient reported outcome* or questionnaire* or PROMS):ti,ab,kw (Word variations have been searched)

Supplementary Table 1: Search strategy per database.

Study	Random sequence generation	Allocation concealment	Blinding of participants	Blinding of outcome	Incomplete outcome data	Selective reporting	Other bias	A priori hypothesis	Rationale Questionnaire	Cultural validity	Coverage domains	Explicit statement	Baseline compliance	Timing assessment	Reasons missing data	Clinical significance	Results in general	Robust reporting	Test significance	HRQL difference	Exploration missing data	Description population	Inclusion/exclusion criteria	Primary/secondary outcomes	Description treatment reported	Initial compliance reported	Follow-up compliance reported	Follow-up compliance ≥ 80%	N≥26 first point	N≥26 all points	Sample size calculation	Control multiple testing	Final assessment
Alderson 2017[57]	+	+	−	+	+	+	+	+	−	+	+	−	+	+	+	+	+	+	+	−	−	+	+	−	+	+	+	−	+	+	+	−	High
Avery 2007[68]	N	N	N	N	−	+	+	−	−	+	+	+	+	?	+	+	+	+	+	+	−	+	+	+	+	+	+	+	+	+	−	−	High
Barbour 2008[69]	N	N	N	N	+	+	+	−	−	+	−	−	+	+	+	+	+	+	−	+	−	+	+	+	+	+	+	−	−	−	−	Low
Barbour 2017[94]	N	N	N	N	?	+	+	+	+	+	+	+	+	+	−	−	+	+	+	+	−	+	+	+	+	+	+	−	+	−	−	High
Blazeby 2003[70]	N	N	N	N	+	−	?	+	−	+	+	+	+	+	+	+	+	−	+	+	−	−	+	+	−	+	+	−	+	−	+	Low
Blazeby 2004[74]	N	N	N	N	+	+	+	−	+	+	+	−	+	+	−	−	+	+	+	+	−	+	+	+	+	+	+	−	+	−	−	High
Blazeby 2005[71]	N	N	N	N	+	−	+	+	+	+	+	−	+	+	+	−	+	−	+	−	−	+	+	+	+	+	−	−	−	−	+	Low
Bonnetain 2006[72]	?	+	?	+	?	+	+	−	+	+	−	+	+	+	−	−	+	+	+	+	−	+	+	+	+	+	−	−	+	−	−	Low
Brooks 2002[73]	N	N	N	N	+	+	+	−	+	+	+	+	+	+	+	−	+	+	+	−	+	+	+	+	−	+	?	−	−	−	−	Low
Cavallin 2015[37]	N	N	N	N	−	−	+	+	+	+	+	+	+	+	−	+	+	−	+	+	+	+	+	+	+	+	?	?	−	−	−	Low
Cense 2006[38]	N	N	N	N	−	−	−	+	−	+	−	−	+	+	−	+	+	−	+	+	−	−	+	+	+	+	+	−	+	−	−	Low
Chou 2010[39]	N	N	N	N	?	−	−	−	−	+	+	−	+	−	+	+	+	−	+	−	−	+	+	+	+	+	+	+	−	−	−	Low
De Boer 2004[40]	N	N	N	N	+	−	−	−	+	+	−	−	+	+	−	+	+	+	+	−	−	+	+	+	+	+	+	+	+	−	−	Low
Egberts 2008[41]	N	N	N	N	+	+	?	+	+	?	+	−	+	+	+	+	+	+	+	−	−	−	+	+	+	+	+	−	+	+	−	−	High
Fang 2004[42]	N	N	N	N	+	+	+	−	−	+	−	+	+	?	−	+	+	+	+	N	−	+	+	+	+	+	+	+	+	+	−	−	High
Garant 2019[43]	N	N	N	N	+	−	−	+	+	+	+	−	−	−	−	+	+	−	+	+	−	+	+	+	+	?	?	?	+	+	−	−	Low
Gillham 2008[44]	N	N	N	N	+	+	+	−	+	+	+	+	+	+	+	−	+	−	+	N	−	+	+	+	+	+	+	−	+	−	−	−	Low
Gradauskas 2006[45]	N	N	N	N	−	−	+	−	−	+	−	−	+	+	+	+	+	−	+	−	−	+	+	+	+	+	−	N	−	−	−	+	Low
Haj Mohammad 2016[46]	N	N	N	N	+	+	+	+	−	+	+	+	+	+	+	+	+	+	+	N	+	+	+	+	+	+	+	+	+	+	−	+	High
Hauser 2015[47]	N	N	N	N	+	−	+	+	−	+	+	−	+	+	+	+	+	−	+	+	−	+	+	+	−	+	+	−	+	+	−	−	Low
Huang 2015[48]	N	N	N	N	−	?	−	+	+	+	+	−	+	+	+	?	+	−	+	N	−	+	+	+	+	+	−	−	+	+	−	−	Low
Kachnic 2011[49]	?	+	?	+	+	−	+	−	+	+	−	−	+	+	+	−	+	+	+	+	+	−	+	+	+	+	+	−	+	+	−	−	Low
Kassam 2010[75]	N	N	N	N	?	+	+	+	+	+	−	−	+	?	+	+	+	−	+	+	−	+	+	+	+	+	+	−	+	−	−	Low
Kim 2012[76]	N	N	N	N	?	+	+	+	−	+	+	+	+	+	−	+	+	+	+	+	−	+	+	+	+	+	+	+	+	+	−	−	High
Kobayashi 2011[77]	N	N	N	N	−	+	−	−	+	+	+	+	+	+	+	+	+	?	+	+	−	+	+	+	+	+	−	+	+	+	−	−	Low
Kong 2012[78]	N	N	N	N	−	+	+	−	−	+	+	+	−	+	−	−	+	+	+	N	−	+	+	+	−	−	−	?	?	?	−	−	Low
Li 2018[79]	+	+	?	+	?	−	+	−	+	+	+	−	+	+	−	−	+	+	+	+	−	+	+	+	+	+	−	?	+	?	−	−	Low
Lv 2014[50]	N	N	N	N	?	+	+	−	−	+	−	−	?	+	−	−	+	−	+	+	−	+	?	+	+	?	−	?	?	?	−	−	Low
Malström 2015[51]	N	N	N	N	+	+	+	−	−	+	+	−	+	+	−	−	+	?	+	+	−	+	+	+	−	+	+	−	+	+	−	−	Low
Nafteux 2013[52]	N	N	N	N	−	−	+	−	−	+	+	+	+	+	+	+	+	?	−	+	−	+	+	+	+	+	+	+	+	+	−	−	Low
Noordman 2018[17,36]	+	+	?	+	+	+	+	+	+	+	+	−	+	+	+	+	+	+	+	+	−	+	+	+	+	+	+	+	+	+	−	+	High
Parameswaran 2010[53]	N	N	N	N	+	−	+	−	−	+	+	−	+	+	−	+	+	+	−	N	−	+	+	+	+	+	+	−	+	+	−	−	High
Park 2018[80]	N	N	N	N	−	+	−	−	−	+	+	−	+	+	−	−	+	?	+	+	−	+	+	+	+	+	+	+	+	+	−	−	Low
Pruthi 2018[81]	N	N	N	N	−	?	−	−	+	+	+	−	−	?	+	+	+	−	+	N	−	+	+	+	+	−	−	N	+	?	−	−	Low

Supplementary Table 2 – *continued from previous page*

Study	Random sequence generation	Allocation concealment	Blinding of partcipants	Blinding of outcome	Incomplete outcome data	Selective reporting	Other bias	A priori hypothesis	Rationale Questionnaire	Cultural validity	Coverage domains	Explicit statement	Baseline compliance	Timing assessment	Reasons missing data	Clinical significance	Results in general	Robust reporting	Test significance	HRQL difference	Exploration missing data	Description population	Inclusion/exclusion criteria	Primary/secondary outcomes	Description treatment reported	Initial compliance reported	Follow-up compliance reported	Follow-up compliance $\geq$ 80%	N$\geq$26 first point	N$\geq$26 all points	Sample size calculation	Control multiple testing	Final assessment
Ramakrishnaiah 2014[54]	N	N	N	N	–	+	–	–	–	+	+	–	?	+	–	–	+	–	+	–	–	+	?	+	+	?	?	?	–	–	–	–	Low
Rees 2015[55]	+	+	?	+	+	+	+	–	+	+	+	–	+	+	+	+	+	+	–	+	–	+	+	+	+	+	+	–	+	+	–	–	High
Reynolds 2006[56]	N	N	N	N	+	–	+	–	+	+	+	–	+	+	+	+	+	–	+	+	–	–	+	+	–	+	+	–	+	+	–	–	Low
Safieddine 2009[18]	N	N	N	N	+	+	+	–	+	+	+	–	+	+	+	+	+	+	+	N	–	+	+	+	+	+	+	–	+	+	–	+	High
Sarkaria 2019[58]	N	N	N	N	+	+	+	+	+	+	+	+	+	+	–	–	+	+	+	+	–	+	+	+	+	+	+	+	+	+	–	–	High
Sugawara 2019[59]	N	N	N	N	+	–	+	–	–	+	+	–	+	+	+	+	+	+	+	+	–	+	?	+	+	+	–	–	+	?	–	–	High
Sunde 2019[60]	N	N	N	N	+	+	+	+	+	+	+	–	+	+	–	+	+	+	+	+	–	+	+	+	+	+	+	–	+	–	–	+	High
Tatematsu 2013a[61]	N	N	N	N	–	+	–	+	+	+	–	–	–	–	N	–	+	–	+	N	–	+	+	+	–	–	–	N	+	+	+	–	Low
Tatematsu 2013b[62]	N	N	N	N	–	–	–	+	–	+	–	–	?	+	–	–	+	–	+	N	–	+	+	+	–	?	?	?	+	?	+	–	Low
van der Sluis 2019[63]	+	+	+	+	–	–	+	–	–	+	+	–	?	+	–	+	+	–	+	+	–	+	+	+	+	–	–	?	?	?	+	–	Low
Van Heijl 2010[64]	N	N	N	N	+	+	+	+	+	+	–	–	+	+	+	+	+	+	+	N	+	+	+	+	+	+	+	+	+	+	–	–	High
Van Meerten 2008[65]	N	N	N	N	+	+	+	–	–	+	+	+	+	+	+	+	+	+	+	N	–	+	+	+	+	+	+	+	+	+	–	–	High
Wu 2008[82]	?	+	?	N	+	+	+	+	–	?	+	+	+	+	–	–	+	+	+	+	–	–	+	+	+	+	+	–	+	+	–	–	High
Yamashita 2014[66]	N	N	N	N	+	+	+	–	+	+	+	+	+	+	–	+	+	+	+	N	–	+	+	+	+	+	+	–	+	+	–	–	High
Zeng 2012[67]	–	–	?	+	–	–	–	–	+	+	+	+	–	+	–	–	+	–	+	+	–	+	+	+	–	–	–	?	?	?	–	–	Low

Supplementary Table 2: Quality assessment of all included studies.
N: not applicable; +: sufficient; –: insufficient; ?: unclear.

Questionnaire	Subscale	Comparison	t-value	p-value
EORTC QLQ-C30	Physical functioning	Neoadjuvant + surgery vs. Definitive Treatment	3.11	0.0038
EORTC QLQ-C30	Physical functioning	Surgery vs. Definitive Treatment	-0.67	0.50
EORTC QLQ-C30	Physical functioning	Neoadjuvant + surgery vs. Surgery	3.85	0.0003
EORTC QLQ-C30	Global health status	Neoadjuvant + surgery vs. Definitive Treatment	3.45	0.0017
EORTC QLQ-C30	Global health status	Surgery vs. Definitive Treatment	0.92	0.36
EORTC QLQ-C30	Global health status	Neoadjuvant + surgery vs. Surgery	3.4	0.0017
EORTC QLQ-C30	Fatigue	Neoadjuvant + surgery vs. Definitive Treatment	-2.46	0.0421
EORTC QLQ-C30	Fatigue	Surgery vs. Definitive Treatment	-1.81	0.14
EORTC QLQ-C30	Fatigue	Neoadjuvant + surgery vs. Surgery	-1.15	0.25
EORTC QLQ-C30	Appetite	Neoadjuvant + surgery vs. Definitive Treatment	-0.31	0.76
EORTC QLQ-C30	Appetite	Surgery vs. Definitive Treatment	-1.83	0.13
EORTC QLQ-C30	Appetite	Neoadjuvant + surgery vs. Surgery	2.66	0.0238
EORTC QLQ-C30	Diarrhoea	Neoadjuvant + surgery vs. Definitive Treatment	-2.58	0.0297
EORTC QLQ-C30	Diarrhoea	Surgery vs. Definitive Treatment	-1.5	0.13
EORTC QLQ-C30	Diarrhoea	Neoadjuvant + surgery vs. Surgery	-1.86	0.13
EORTC QLQ-C30	Role functioning	Neoadjuvant + surgery vs. Definitive Treatment	2.56	0.0210
EORTC QLQ-C30	Role functioning	Surgery vs. Definitive Treatment	-0.38	0.70
EORTC QLQ-C30	Role functioning	Neoadjuvant + surgery vs. Surgery	3.49	0.0015
EORTC QLQ-C30	Cognitive functioning	Neoadjuvant + surgery vs. Definitive Treatment	2.48	0.0397
EORTC QLQ-C30	Cognitive functioning	Surgery vs. Definitive Treatment	1.03	0.30
EORTC QLQ-C30	Cognitive functioning	Neoadjuvant + surgery vs. Surgery	1.97	0.10
EORTC QLQ-C30	Financial worry	Surgery vs. Neoadjuvant + surgery	-2.35	0.0188
EORTC QLQ-OES	Taste	Neoadjuvant + surgery vs. Definitive Treatment	-4.8	< 0.0001
EORTC QLQ-OES	Taste	Surgery vs. Definitive Treatment	-6.45	< 0.0001
EORTC QLQ-OES	Taste	Neoadjuvant + surgery vs. Surgery	-0.12	0.90
EORTC QLQ-OES	Dysphagia	Neoadjuvant + surgery vs. Definitive Treatment	-4.54	< 0.0001
EORTC QLQ-OES	Dysphagia	Surgery vs. Definitive Treatment	-3.84	0.0002
EORTC QLQ-OES	Dysphagia	Neoadjuvant + surgery vs. Surgery	-1.58	0.11
EORTC QLQ-OES	Reflux	Surgery vs. Neoadjuvant + surgery	2.94	0.0032
EORTC QLQ-OES	Pain	Neoadjuvant + surgery vs. Definitive Treatment	-1	0.32
EORTC QLQ-OES	Pain	Surgery vs. Definitive Treatment	-3.19	0.0043
EORTC QLQ-OES	Pain	Neoadjuvant + surgery vs. Surgery	2.05	0.08
EORTC QLQ-OES	Choking while swallowing	Surgery vs. Neoadjuvant + surgery	1.98	0.0481
EORTC QLQ-OES	Speech	Surgery vs. Neoadjuvant + surgery	5.72	< 0.0001

Supplementary Table 3: One-sided post-hoc tests in analyses between treatment groups neoadjuvant treatment, definitive treatment, and surgery in the short-term analysis comparing the effect of treatment on different HRQoL subscales.

Chapter 9

Prediction of health-related quality of life in patients with potentially curable oesophageal or gastric cancer

Héctor G. van den Boorn, Steven C. Kuijper, Jessy Joy van Kleef, Rob H.A. Verhoeven, Suzanne S. Gisbertz, Walther Jansen, Jelle P. Ruurda, Camiel Rosman, Marije Slingerland, Aeilko H. Zwinderman, Mirjam A.G. Sprangers, Hanneke W.M. van Laarhoven and Ameen Abu-Hanna

Submitted

Abstract

Introduction

Curative treatment of oesophageal and gastric cancer often comes at the cost of health-related quality of life (HRQoL). Models predicting HRQoL could aid in the process of shared decision making by informing patients how their HRQoL could be affected by treatment. However, currently no models to predict HRQoL exist for oesophagogastric cancer patients. The aim of this study was to develop a prediction model that predicts HRQoL in individual patients three and six months after the start of treatment.

Methods

The data of potentially curable oeosphagogastric cancer patients (with TNM staging of $cT_{1\text{-}4a,X}N_{0\text{-}3,X}M_0$) was obtained from the POCOP registry and The Netherlands Cancer Registry. Prediction models were developed to predict the EORTC summary score at three ($N = 537$) and six months ($N = 480$) after the start of the treatment. The model predictors included patient, tumour and treatment characteristics. The regression models were developed using the Extreme Gradient Boosting (XGboost) algorithm. The models were internally validated through nested cross-validation, using the root mean squared error (RMSE) as performance measure.

Results

In both analyses, there was a large variety in the change of HRQoL, with around 35% of patients experiencing a clinically relevant change in summary score. Baseline summary score was found to be predictive of the summary score at the end of the period ($p < 0.001$) in both analyses. WHO performance status was also associated with the summary score at six months ($p = 0.017$). Compared to a non-informative model, the RMSE was reduced from 15.3 to 12.9 (16% relative reduction) in the three-months analysis ($p < 0.001$). In the six-months analysis, the RMSE was reduced from 14.2 to 12.6 (11% relative reduction; $p < 0.001$).

Discussion

Although the prediction models outperformed non-informative models, the overall reduction in RMSE and predictive performance was relatively low for both models, which limits the added value of the models in clinical practice. The study demonstrates the importance of incorporating additional predictive variables in order to improve the predictive performance of the models.

Introduction

The prognosis in patients with oesophageal and gastric cancer is poor and even after treatment with curative intent, the five-year overall survival rate often is less than 50%.[1–4] In the curative setting treatment options include, among others, perioperative chemotherapy, definitive chemoradiation, (neo)adjuvant chemo(radio)therapy and resection alone.[5–10] The curative treatments often result in an increased overall survival, however this can come at the cost of health-related quality of life (HRQoL).[5,11–13] Patients who underwent curative surgery often indicate that enhancing or maintaining their HRQoL is one of their primary objectives.[14] Consequently, in deciding between cancer treatment options, HRQoL of patients is increasingly recognised as an important outcome that needs to be considered.[15] In the context of shared decision making, it is therefore important to have a realistic view of the expected HRQoL, as this may influence the eventual choice of treatment.[16]
Recently, a systematic review and meta-analysis investigated the influence of treatment on HRQoL in curatively treated oesophagogastric cancer patients[13]. Results included average HRQoL patterns in large populations, however, average HRQoL patterns are of limited value in shared decision making because HRQoL in individual patients may vary substantially. Prediction models that predict HRQoL based on patient, tumour and treatment characteristics could therefore aid physicians and patients in providing personalised and accurate information. Such models predicting HRQoL are absent within oesophagogastric oncology.[17] However, in other domains of oncology, few models are available that predict HRQoL in individual patients. One example is the PrediQt-Cx model for cervical cancer patients.[18] This model predicts a dichotomised HRQoL (low risk versus high risk of a low level of HRQoL) six months after treatment using a support vector machine. Another example is a model predicting one year risk of having a low level of HRQoL for patients suffering from colorectal cancer.[19] This latter study used logistic regression to model the risk of poor HRQoL. Although these prediction models may have promising results, such models are not generalisable to other populations of cancer patients.

The aim of this study is therefore to develop and internally validate prediction models that predict HRQoL in individual patients prior to the start of treatment. These prediction models are meant to offer patients a quantitative assessment of how their HRQoL is likely to be affected in the short term due to treatment.

Methods

This report is written according to the TRIPOD guidelines (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis).[20] The data used in this analysis was obtained from two sources, the POCOP registry[21] and the Netherlands Cancer Registry[22]. As the aim of this study is to detect patterns in HRQoL following treatment in the short term, this analysis focuses on two end-points: at three months after start of treatment, when a large proportion of patients has not undergone resection yet, and at six months after the start of treatment when the resection has taken place for most patients.

The population-based POCOP registry is a database containing patient-reported outcome measures for HRQoL in patients with oesophagogastric cancer.[21] All patients treated in collaborating centres are invited to participate. The data of the POCOP registry was linked to the prospectively maintained Netherlands Cancer Registry (NCR)[22], a nation-wide database containing patient, tumour and treatment characteristics. Participants are asked to fill in questionnaires at inclusion (diagnosis) and every three months during the first year, and (bi-)annually thereafter. One of the included measures is the EORTC QLQ-C30[23], a widely employed and validated 30-item instrument to measure HRQoL, whose responses were used in this analysis. The questionnaire includes five function scales (physical, role, cognitive, emotion and social functioning), a global health status scale, three symptom scales (nausea/vomiting, fatigue and pain) and six single items (appetite loss, diarrhoea, dyspnoea, constipation, insomnia and financial difficulties).[24] Subsequently, the summary score was calculated, following the EORTC guidelines, as the average of 13 QLQ-C30 outcome scores, excluding global health status and financial difficulties.[25] The summary score was based on all completed items, and was set to missing if over six of the necessary scores were missing.[25] The summary score ranges from 0–100, with a higher scores indicating better HRQoL.[26] In order to determine and interpret the magnitude of the clinically relevant change in summary score, we used the threshold values and cut-off scores determined by Cocks et al.[26]. As clinically meaningful threshold values for the summary score have not been reported yet, we calculated this threshold as the median clinically meaningful differences for the scales on which the summary score was based. A clinically meaningful change in the summary score was therefore determined to be 9 points for an increase in summary score and 11 points for a decrease in summary score.

The QLQ-C30 data of all 898 potentially curable patients (with TNM staging of $cT_{1\text{-}4a,X}N_{0\text{-}3,X}M_0$)[4] diagnosed between 2015 and 2020 were retrieved from the POCOP registry. Patients with missing summary scores were excluded from the analysis. As the aim of this study is to analyse the effects of treatment on HRQoL, patients must have completed at least two questionnaires over time. The baseline questionnaire must be completed before or within 14 days after the start of treatment. As the analysed patterns span the first three and the first six months following treatment's start, patients were excluded if they completed their last questionnaire 14 days before the end of this period or earlier. A patient inclusion flowchart is provided in Figure 1. An additional analysis was performed to determine whether patients who were excluded due to missing questionnaires, differed from patients whose data were included in this study. In this exclusion analysis, patient, tumour and treatment characteristics are compared using ANOVA (for numerical variables) or chi-squared tests (for categorical variables) with Holm's significance correction for multiple testing. A Kaplan-Meier analysis was also performed to determine differences in overall survival between included and excluded patients.

Prediction of summary score

In order to predict the summary score at three and at six months following treatment, two prediction models were developed based on a priori available information at the time of diagnosis. In this case patient characteristics (age, sex, body mass index, WHO performance status, education level, marital status, weight loss, smoking and alcohol use, hemoglobine level, creatinine and LDH), tumour characteristics (location of tumour and cTNM staging), treatment type (see Table 1), and baseline summary score were used as predictors. Missing predictor data were handled with multiple imputations with five iterations by chained equations.[27] The Extreme Gradient Boosting (XGBoost) algorithm[28] was used to predict the summary score at the end of the period given the predictors. This algorithm has previously been used successfully in various settings and generally has a high performance.[29–31] To account for the variability due to the choice of the training and test set, hence to increase robustness and generalisability, a 5x2-fold nested cross validation was employed. With this internal validation scheme, the data is randomly split into two equal parts; one part is used to train the prediction model, and the other is used for model validation and vice versa. This process is repeated five times to establish the variability of the performance.[32]

Figure 1: Patient inclusion flowchart for the analyses at three and six months following the start of treatment.

The XGBoost algorithm is an ensemble learner which utilises multiple regression trees in an ensemble to create a single prediction model that can outperform the individual regression trees. Predictions are made by utilising voting on the individual trees, where each vote is weighed by the performance of that regression tree. XGBoost depends on a number of so-called 'hyper-parameters'; parameters that are not estimated directly from the dataset but must be given a priori. The hyper-parameters on which XGBoost depends are the learning rate (η), minimum loss reduction required to continue partitioning of leaf nodes (γ), the maximum depth of the regression trees, sub-

sample ratio of variables and the number of boosting iterations. To determine which realisation of hyper-parameters leads to the most optimal classifier, a nested round of 10-fold cross-validation[33] was performed. In the inner cross-validation, hyper-parameter combinations are used for training and evaluated on the corresponding hold-out dataset. The combination of hyper-parameters that provided the best overall root mean squared error (RMSE)[34] in the inner cross-validation was then used to train a model on the whole training set which is then evaluated on the independent outer test set. This process is illustrated in Figure 2. All analyses were performed in the RStudio environment with R version 4.0.3 using the *xgboost* package.[28,35]

Model evaluation

The models were evaluated using the RMSE between the observed and the predicted summary scores at three and six months. In order to evaluate the magnitude of this score, the models' RMSE was compared to the RMSE of a non-informative model, which uses the observed mean summary score as a constant prediction for all patients. A Wilcoxon matched-pairs (signed-rank) test was used to evaluate if the RMSEs of the prediction models in the five repeated data splits were lower than the RMSE of the non-informative models. This is a necessary but not sufficient property of a clinically useful prediction model. Furthermore, the association of the predictors with the summary score at three and six months were evaluated. An ANOVA was used to evaluate the associations with categorical variables, and Spearman correlation coefficients were calculated for continuous predictors. All p-values were corrected for multiple testing using the Holm-Bonferroni method to reduce the risk of Type-I errors.

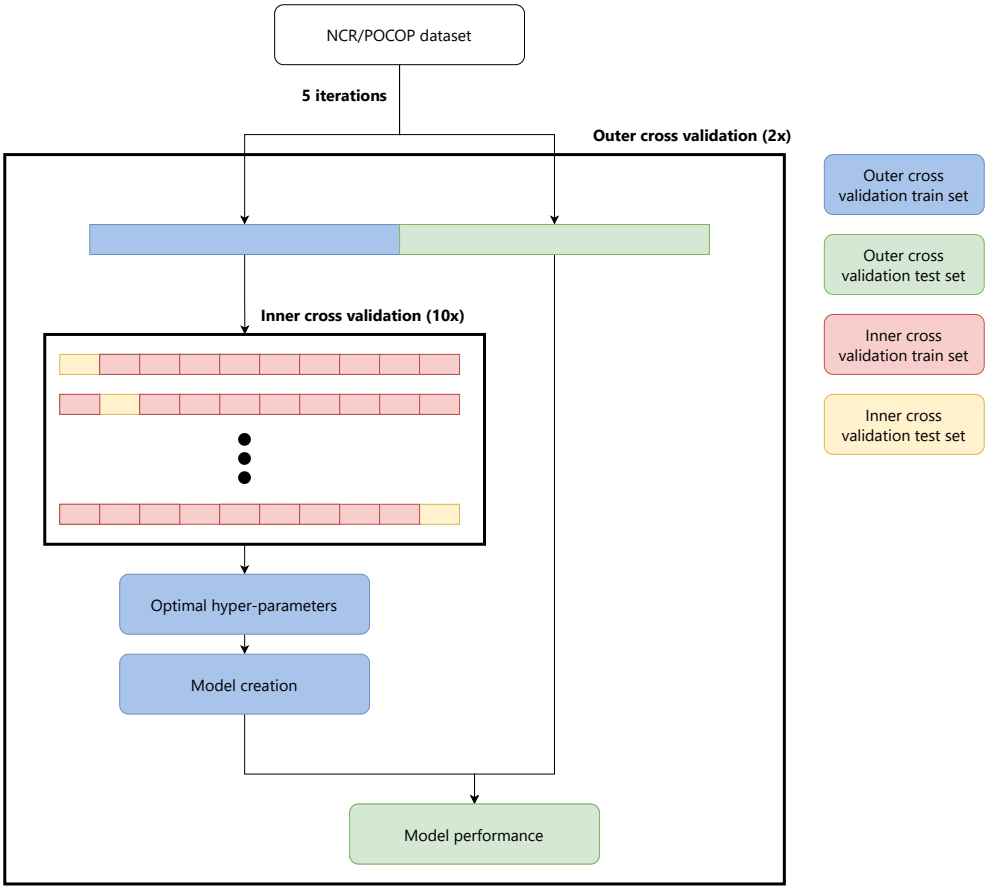


Figure 2: Model creation and validation.
This figure illustrates the 5x2 nested cross validation mechanism which was used to optimise hyper-parameters and to evaluate the model on test sets. For both the three months and six-months analyses, the NCR/POCOP dataset was split into two folds of equal size. One part, the train set, in the outer cross validation (shown in blue) was used to create a prediction model; the test set (shown in green) consists of the remaining data and was used to evaluate the performance of the prediction model. This process was repeated so that every patient in the dataset was used exactly once for model evaluation. The outer train set (shown in blue) was used for model creation; the optimal hyper-parameters were determined with an additional layer of cross validation. Within this inner cross validation layer, all combinations of hyper-parameters were tested and evaluated on the inner cross validation test set (shown in yellow) to establish the optimal combination of hyper-parameters. These settings were used to create a prediction model with the outer train set (blue) and model performance was determined on the outer test set (in green). To evaluate the variability of the models' performance, the entire train-test validation scheme was repeated five times.[32]

Results

An overview of the main patient characteristics is given in Table 1 for the included patients. An overview of additional variables used in the analyses is given in Supplementary Table 1.

A total of 537 patients were included in the three-months analysis out of a total of 898 patients. A subset of these 537 patients completed questionnaires until the end of the six-month period and were also included in the six-months analysis (N = 480 patients). Of all patients who underwent surgery, a total of 48% did so within three months of the start of treatment, and 99% of patients within six months. Most of the excluded patients did not complete the questionnaire before the start of treatment (N = 270) or did not fill in the questionnaire at the end of the study period (respectively N = 78 and N = 135). Patients who were excluded because the first questionnaire was filled out too late, did not differ from included patients in terms of patient, tumour and treatment characteristics or overall survival (Supplementary Tables 2 and 3). However, patients who were lost to follow-up and therefore excluded from the analysis, differed in cT staging of the tumour and were found to be over five times as likely to die compared to the included patients. Furthermore, patients who were lost to follow-up also differed in received treatment, were more likely to have undergone chemoradiation (with low dose radiation) and less likely to receive neoadjuvant chemoradiation ($p < 0.001$).

In the three-months analysis, the summary score at diagnosis was 83.4 (SD: 12.6), and at the end of the period was 79.2 (SD: 15.5). In the six-months analysis, the summary score changed from 83.8 (SD: 12.8) at baseline to 77.7 (SD: 14.4) by the end of the analysis The change in summary score is displayed in Figure 3 for both analyses. A clinically meaningful change in summary score was observed in a total of 178 patients (33.1%) in the three-months analysis, and in 186 patients (38.8%) for the six-months analysis.

Variable	Three months - N (%)	Six months - N (%)
N	537	480
Age, mean (SD)	66.07 (8.46)	66.03 (8.41)
Sex		
Female	120 (22.3)	100 (20.8)
Male	417 (77.7)	380 (79.2)
BMI, mean (SD)	26.08 (4.34)	26.02 (4.39)
Missing	77 (14.3)	67 (14.0)
WHO performance status		
Missing	91 (16.9)	80 (16.7)
0	274 (51.0)	248 (51.7)
1	153 (28.5)	136 (28.3)
2	17 (3.2)	14 (2.9)
3+	2 (0.4)	2 (0.4)
Tumour location		
Oesophagus	464 (86.4)	411 (85.6)
Stomach	73 (13.6)	69 (14.4)
cT stage		
cT_1	9 (1.7)	8 (1.7)
cT_2	182 (33.9)	166 (34.6)
cT_3	297 (55.3)	266 (55.4)
cT_4	9 (1.7)	6 (1.2)
cT_X	40 (7.4)	34 (7.1)
cN stage		
cN_0	264 (49.2)	244 (50.8)
cN_1	180 (33.5)	162 (33.8)
cN_2	86 (16.0)	67 (14.0)
cN_3	7 (1.3)	7 (1.5)
Treatment		
Chemoradiation (high dose)	48 (8.9)	41 (8.5)
Chemoradiation (low dose)	35 (6.5)	27 (5.6)
Neoadjuvant chemoradiation and surgery	324 (60.3)	296 (61.7)
Neoadjuvant chemotherapy and surgery	37 (6.9)	31 (6.5)
Perioperative chemotherapy	16 (3.0)	9 (1.9)
Resection (primary tumour)	49 (9.1)	49 (10.2)
Other	28 (5.2)	27 (5.6)
Resection performed by end of period		
No	226 (42.1)	4 (0.8)
No resection	99 (18.4)	77 (16.0)
Yes	212 (39.5)	399 (83.1)
Number of questionnaires		
2	235 (43.8)	38 (7.9)
3	301 (56.1)	223 (46.5)
4	1 (0.2)	217 (45.2)
5	0 (0.0)	2 (0.4)

Table 1: Characteristics overview of included patients.
cT and cN stages are defined according to the 8th edition of the UICC Cancer Staging Manual. Tumour location is defined according to the ICD-O-3. SD: standard deviation.

Model performance

The non-informative models, using the mean summary score as a constant prediction, resulted in an RMSE of 15.3 (95% CI: 15.1–15.5) in the three-months analysis, and an RMSE of 14.2 (95% CI: 13.8–14.7) in the six-months analysis. The nested cross-validated RMSE of the prediction model predicting the summary score at three months was 12.9 (95% CI: 12.6–13.2) which was lower than the non-informative RMSE ($p < 0.001$) with an improvement of 15.7%. For the six-months analysis, the RMSE of the prediction model was 12.6 (95% CI: 12.2–12.9) which was also lower than the corresponding non-informative RMSE ($p < 0.001$) with an improvement of 11.3%.

Predictor associations

Supplementary Table 4 displays the associations of the predictors with the summary score at respectively three and six months. For the summary score at three months, only baseline summary score was found to be predictive of the outcome ($\rho = 0.583$, $p < 0.001$). In the six-months analysis, baseline summary score was also found to be predictive of the outcome ($\rho = 0.546$, $p < 0.001$), as well as WHO performance status ($p = 0.017$). Subsequent paired Wilcoxon post-hoc tests using Holm-Bonferroni correction indicated that summary score differed only between performance status 0 and 2 ($p = 0.032$).

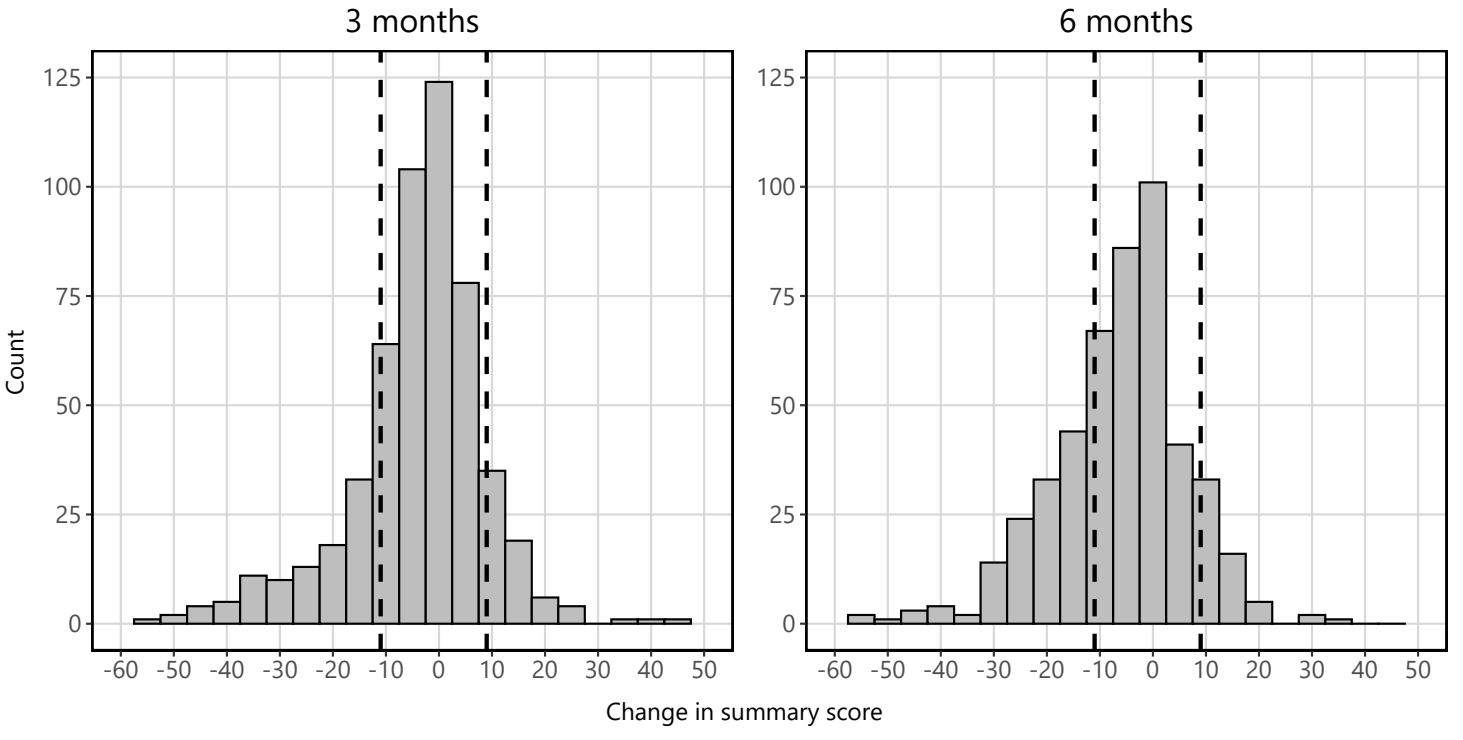


Figure 3: Change in summary score between baseline questionnaire and the end of the analysis period.
A positive change in summary score indicates an improvement of summary score at the end of the analysis period compared to the baseline measurement. The vertical dashed lines indicate the boundaries of clinically significant change in summary score (at -11 and 9 points).

Discussion

In this study we developed and evaluated prediction models predicting the EORTC summary score at three and six months in potentially curable oesophagogastric patients. In both models, a decrease in the root mean squared error (RMSE) compared to a non-informative model was observed, though this decrease was modest and improved upon the non-informative prediction by only 16% and 11% for three and six months, respectively.

When comparing the mean summary score at the end of the analysis period to the baseline scores, only a small difference is observed. However, large variations in the changes of summary score are observed across individual patients, as indicated in Figure 3, and in approximately 35% of patients this change can be considered to be clinically meaningful. The goal of this study was to predict these individual differences and to offer patients a quantitative prediction of the expected HRQoL. The models demonstrated only a small decrease in RMSE and therefore only accounted for a small part of the variability in the summary score, which limits their application in clinical practice. The low performance may be explained by the limited number of predictive variables that were found. Only the baseline summary score and the WHO performance status were found to be associated with the summary score at the end of the analysis period. The significant effect of baseline summary with the summary score at the time-point might be an artificial effect due to regression toward the mean. This would imply that patients with very low or high baseline summary scores will, on average, have respectively higher or lower summary scores at the end of the analysis period. In contrast to our study, previous studies that have investigated change in HRQoL over time in cancer patients found more variables that were predictive such as: age[36,37], education level[36], body mass index[38] and treatment[38,39]. The approach in the current study differs from these studies and HRQoL was analysed during the first three and six months after treatment, whereas previous studies reported results in the long term; i.e. 12 months[37], 5.5 years[36], 6 years[38] and 6.5 years[39]. In this study, we showed that variables such as age, body mass index, treatment and education level may not be associated with HRQoL in the short term.

We hypothesise that our model can further be improved by identifying and using variables that are more predictive of HRQoL. One of those variables might be the illness perceptions of patients. Prior studies have found that illness perceptions of cancer patients can be predictive of future reported HRQoL.[40]

Future research could investigate the prognostic value of illness perceptions in oesophagogastric cancer patients to predict HRQoL in the short term after the treatment, to increase the performance of HRQoL prediction models. Additionally, the inclusion of more patients in the POCOP study can lead to prediction models with more fine-tuned parameters, a smaller potential for model overfitting, and overall better predictions.[41]

Limitations and strengths

The current study has a number of limitations. First, participation in the POCOP study is voluntary, and therefore, patient self-selection can introduce a potential source of bias in the relationship between the outcome and exposure.[42] For example, the emotional, cognitive and or physical condition of the patients might prevent from or attract patients to participating in the study.[43] In addition, self-administered HRQoL questionnaires have been found to induce bias against patients with comorbidities.[44] The low overall survival of patients who were lost to follow-up in our study could be due to a high level of comorbidities and/or impaired physical condition.

Furthermore, the overall generalisability of the results may be limited. There was a substantial part of the cohort data that could not be used for the analysis. However, patients who were excluded because their first questionnaire was completed after the start of treatment, were found not to be different from included patients in terms of patient, tumour or treatment characteristics or overall survival. Their exclusion was therefore not likely to influence the results. Patients who were lost to follow-up before the end of the analysis period, however, did differ from included patients in terms of cT staging, treatment and overall survival. In this case, patients may have been lost to follow-up due to poor HRQoL, which could have affected the generalisability of our findings to patients with a low HRQoL during or after treatment.

Another limitation of this study is that the treatment intent could not be determined, because the NCR only reports on treatments that patients actually received. In determining which treatment patients received, no distinction can therefore be made between, for example, patients who were meant to undergo definitive chemoradiation or patients for whom the first choice was neoadjuvant chemoradiation but who did not advance to surgery. This may also explain why patients who were lost to follow-up, were more likely to have received chemoradiation (with low-dose radiation) as a treatment as opposed to neoadjuvant chemoradiation followed by surgery.[45]

Finally, there was variability in the treatment duration among patients. Ideally, patients would fill out the questionnaires at the start of treatment, directly after treatment and a few months thereafter. However, in the POCOP study,

patients filled out the questionnaires at set dates following diagnosis, and patients filling out a certain questionnaire may be in different stages of treatment. This is illustrated by the fact that almost half the patients underwent resection within three months after the start of treatment, and nearly all patients completed their resection within six months after the start of treatment. Due to this limitation, the direct effect of treatment on HRQoL could be measured less accurately.

This study also has several strengths. First, during the development phase of the prediction models, we provided unbiased estimates of model performance through nested cross-validation, giving an estimate of the model's performance on unseen data. Second, by using XGBoost, non-linear relations and interactions in the relationship between the predictors and summary score and predictors are modelled. Third, the effects of bias due to missing data were reduced by handling missing data through multiple imputation. This is a superior method to complete case analysis which has been found to cause bias when the data cannot be assumed to be missing at random.[46] Finally, compared to the models that have previously been developed for HRQoL in cervical[18] and colorectal cancer patients[19], the aim of our approach is to provide patients with quantitative information about their prospective HRQoL. In these prior models the HRQoL outcome was dichotomised into low versus high risk of poor HRQoL, whereas we aimed to predict the summary score itself in order to inform the patient to which degree HRQoL of comparable patients may deteriorate or even improve.

To conclude, prediction models were constructed and evaluated to predict HRQoL outcomes for oesophagogastric cancer patients three and six months after start of the treatment. The use of prediction models that accurately predict HRQoL can provide patients and physicians potentially valuable information, especially if maintaining or increasing HRQoL is a treatment goal. Results from the current nested cross-validations, however, showed that the predictive performance of our models is low, precluding their clinical implementation. It is furthermore unlikely that the models developed in the current study can aid in shared decision making, as treatment was found not to be related to the summary score in the short-term. However, with the identification of more predictive variables in future research, it is our hope that the performance of the models can be improved, and will provide patients with a realistic outlook on their HRQoL.

References

1. Badgwell, B., Das, P. & Ajani, J. Treatment of localized gastric and gastroesophageal adenocarcinoma: The role of accurate staging and preoperative therapy. Journal of Hematology and Oncology 10, 1–7 (2017).
2. IKNL. Nederlandse Kankerregistratie (NKR). `iknl.nl/nkr-cijfers` (2021).
3. Van Hagen, P., Hulshof, M., van Lanschot, J., *et al.* Preoperative Chemoradiotherapy for Esophageal or Junctional Cancer. New England Journal of Medicine 366, 2074–2084 (2012).
4. Van den Boorn, H. G., Abu-Hanna, A., Haj Mohammad, N., *et al.* SOURCE: Prediction models for overall survival in patients with metastatic and potentially curable esophageal and gastric cancer. Journal of Clinical Oncology 38, 301–301 (2020).
5. Al-Batran, S. E., Homann, N., Pauligk, C., *et al.* Perioperative chemotherapy with fluorouracil plus leucovorin, oxaliplatin, and docetaxel versus fluorouracil or capecitabine plus cisplatin and epirubicin for locally advanced, resectable gastric or gastro-oesophageal junction adenocarcinoma (FLOT4): a randomised, phase 2/3 trial. The Lancet 393, 1948–1957 (2019).
6. Ando, N., Kato, H., Igaki, H., *et al.* A randomized trial comparing postoperative adjuvant chemotherapy with cisplatin and 5-fluorouracil versus preoperative chemotherapy for localized advanced squamous cell carcinoma of the thoracic esophagus (JCOG9907). Annals of Surgical Oncology 19, 68–74 (2012).
7. Kitagawa, Y., Uno, T., Oyama, T., *et al.* Esophageal cancer practice guidelines 2017 edited by the Japan esophageal society: part 2. Esophagus 16, 25–43 (2019).
8. Lordick, F., Mariette, C., Haustermans, K., *et al.* Oesophageal cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Annals of Oncology 27, v50–v57 (2016).
9. Smyth, E. C., Verheij, M., Allum, W., *et al.* Gastric cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Annals of Oncology 27, v38–v49 (2016).
10. Hulshof, M. C. & van Laarhoven, H. W. Chemoradiotherapy in tumours of the oesophagus and gastro-oesophageal junction. Best Practice and Research: Clinical Gastroenterology 30, 551–563 (2016).
11. Jacobs, M., Macefield, R. C., Elbers, R. G., *et al.* Meta-analysis shows clinically relevant and long-lasting deterioration in health-related quality of life after esophageal cancer surgery. Quality of Life Research 23, 1097–1115 (2014).
12. Scarpa, M., Valente, S., Alfieri, R., *et al.* Systematic review of health-related quality of life after esophagectomy for esophageal cancer. World Journal of Gastroenterology 17, 4660–4674 (2011).
13. Van den Boorn, H. G., Stroes, C. I., Zwinderman, A. H., *et al.* Health-related quality of life in curatively-treated patients with esophageal or gastric cancer: A systematic review and meta-analysis. Critical Reviews in Oncology/Hematology 154, 103069 (2020).
14. Thrumurthy, S. G., Morris, J. J., Mughal, M. M. & Ward, J. B. Discrete-choice preference comparison between patients and doctors for the surgical management of oesophagogastric cancer. British Journal of Surgery 98, 1124–1131 (2011).
15. Alghamedi, A., Buduhan, G., Tan, L., *et al.* Quality of life assessment in esophagectomy patients. Annals of Translational Medicine 6, 84–84 (2018).
16. Stiggelbout, A. M., Van der Weijden, T., De Wit, M. P., *et al.* Shared decision making: Really putting patients at the centre of healthcare. The BMJ (Online) 344, e256 (2012).

17. Van den Boorn, H. G., Engelhardt, E. G., van Kleef, J., *et al.* Prediction models for patients with esophageal or gastric cancer: A systematic review and meta-analysis. PLoS ONE 13, 1–20 (2018).
18. Kumar, S., Rana, M. L., Verma, K., *et al.* PrediQt-Cx: Post treatment health related quality of life prediction model for cervical cancer patients. PLoS ONE 9, e89851 (2014).
19. Révész, D., Van Kuijk, S. M., Mols, F., *et al.* Development and internal validation of prediction models for colorectal cancer survivors to estimate the 1-year risk of low health-related quality of life in multiple domains. BMC Medical Informatics and Decision Making 20, 1–13 (2020).
20. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. BMC Medicine 13 (2015).
21. Coebergh van den Braak, R. R., van Rijssen, L. B., van Kleef, J. J., *et al.* Nationwide comprehensive gastro-intestinal cancer cohorts: the 3P initiative. Acta Oncologica 57, 195–202 (2018).
22. Netherlands Comprehensive Cancer Organisation. IKNL and the NCR. 2020. `https://www.iknl.nl/en`.
23. Aaronson, N. K., Ahmedzai, S., Bergman, B., *et al.* The European organization for research and treatment of cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. Journal of the National Cancer Institute 85, 365–376 (1993).
24. European Organisation for Research and Treatment of Cancer. The EORTC QLQ-C30 Scoring Manual. Tech. rep. (2001).
25. Giesinger, J. M., Kieffer, J. M., Fayers, P. M., *et al.* Replication and validation of higher order models demonstrated that a summary score for the EORTC QLQ-C30 is robust. Journal of Clinical Epidemiology 69, 79–88 (2016).
26. Cocks, K., King, M. T., Velikova, G., *et al.* Evidence-based guidelines for interpreting change scores for the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. European Journal of Cancer 48, 1713–1721 (2012).
27. Van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. Journal of Statistical Software 45, 1–67 (2011).
28. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. `https://github.com/dmlc/xgboost` (2016).
29. Gumus, M. & Kiran, M. S. Crude oil price forecasting using XGBoost. 2nd International Conference on Computer Science and Engineering, UBMK 2017, 1100–1103 (2017).
30. Ren, X., Guo, H., Li, S., *et al.* A novel image classification method with CNN-XGBoost model. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 10431 LNCS, 378–390 (2017).
31. Torlay, L., Perrone-Bertolotti, M., Thomas, E. & Baciu, M. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. Brain Informatics 4, 159–169 (2017).
32. Dietterich, T. G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation 10, 1895–1923 (1998).
33. Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society: Series B (Methodological) 36, 111–133 (1974).
34. Chai, T. & Draxler, R. R. Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. Geoscientific Model Development 7, 1247–1250 (2014).

35. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2018.
36. Dunn, J., Ng, S. K., Breitbart, W., *et al.* Health-related quality of life and life satisfaction in colorectal cancer survivors: Trajectories of adjustment. Health and Quality of Life Outcomes 11, 1–8 (2013).
37. Park, J. H., Jung, Y. S., Kim, J. Y., *et al.* Trajectories of health-related quality of life in breast cancer patients. Supportive Care in Cancer 28, 3381–3389 (2020).
38. Farris, M. S., Kopciuk, K. A., Courneya, K. S., *et al.* Identification and prediction of health-related quality of life trajectories after a prostate cancer diagnosis. International Journal of Cancer 140, 1517–1527 (2017).
39. Chambers, S. K., Ng, S. K., Baade, P., *et al.* Trajectories of quality of life, life satisfaction, and psychological adjustment after prostate cancer. Psycho-Oncology 26, 1576–1585 (2017).
40. Ashley, L., Marti, J., Jones, H., *et al.* Illness perceptions within 6 months of cancer diagnosis are an independent prospective predictor of health-related quality of life 15 months post-diagnosis. Psycho-Oncology 24, 1463–1470 (2015).
41. Ying, X. An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series 1168, 022022 (2019).
42. Biele, G., Gustavson, K., Czajkowski, N. O., *et al.* Bias from self selection and loss to follow-up in prospective cohort studies. European Journal of Epidemiology 34, 927–938 (2019).
43. Addington-Hall, J. & Kalra, L. Measuring quality of life: Who should measure quality of life? The BMJ 322, 1417–1420 (2001).
44. Unruh, M., Yan, G., Radeva, M., *et al.* Bias in assessment of health-related quality of life in a hemodialysis population: A comparison of self-administered and interviewer-administered surveys in the HEMO Study. Journal of the American Society of Nephrology 14, 2132–2141 (2003).
45. Koëter, M., van Putten, M., Verhoeven, R. H., *et al.* Definitive chemoradiation or surgery in elderly patients with potentially curable esophageal cancer in the Netherlands: a nationwide population-based study on patterns of care and survival. Acta Oncologica 57, 1192–1200 (2018).
46. Knol, M. J., Janssen, K. J. M., Donders, A. R. T., *et al.* Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values : an empirical example. Journal of Clinical Epidemiology 63, 728–736 (2010).

Supplement

Variable	Three months - N (%)	Six months - N (%)
Education level		
Missing	3 (0.6)	3 (0.6)
Primary school education	41 (7.6)	38 (7.9)
Lower vocational education	189 (35.2)	168 (35.0)
Secondary vocational education	171 (31.8)	151 (31.5)
Higher education	133 (24.8)	120 (25.0)
Marital status		
Missing	3 (0.6)	2 (0.4)
Divorced	35 (6.5)	27 (5.6)
Married	426 (79.3)	390 (81.2)
Never married	30 (5.6)	27 (5.6)
Widowed	43 (8.0)	34 (7.1)
Weight loss		
Missing	2 (0.4)	2 (0.4)
Do not know	22 (4.1)	20 (4.2)
No	168 (31.3)	154 (32.1)
Weight gain	18 (3.4)	17 (3.5)
Yes, desired	52 (9.7)	48 (10.0)
Yes, undesirably	275 (51.2)	239 (49.8)
Smoking status		
No	487 (90.7)	436 (90.8)
Yes	50 (9.3)	44 (9.2)
Alcohol use		
Missing	7 (1.3)	6 (1.2)
No, but consumed alcohol in past	163 (30.4)	141 (29.4)
No, never	40 (7.4)	36 (7.5)
Yes	327 (60.9)	297 (61.9)
Hemoglobine, mean (SD)	8.56 (1.20)	8.56 (1.21)
Missing	43 (8.0)	38 (7.9)
Creatinine, mean (SD)	83.03 (24.81)	82.80 (21.22)
Missing	81 (15.1)	69 (14.4)
Lactate dehydrogenase, mean (SD)	188.17 (43.60)	186.90 (40.51)
Missing	103 (19.2)	89 (18.5)

Supplementary Table 1: Additional characteristics overview of included patients. cT and cN stages are defined according to the 8th edition of the UICC Cancer Staging Manual. Tumour location is defined according to the ICD-O-3. SD: standard deviation.

	Included (N = 537)	Late start (N = 270)	p-value	Lost to follow-up (N = 78)	p-value
Age			0.684		0.587
Mean (SD)	66.1 (8.5)	65.0 (8.7)		67.8 (8.9)	
Sex			1.000		1.000
Female	120 (22.3%)	50 (18.5%)		20 (25.6%)	
Male	417 (77.7%)	220 (81.5%)		58 (74.4%)	
BMI			1.000		1.000
Mean (SD)	26.1 (4.3)	26.3 (4.2)		25.7 (4.0)	
WHO performance status			1.000		0.587
0	274 (61.4%)	136 (55.7%)		39 (65.0%)	
1	153 (34.3%)	96 (39.3%)		17 (28.3%)	
2	17 (3.8%)	10 (4.1%)		2 (3.3%)	
3+	2 (0.4%)	2 (0.8%)		2 (3.3%)	
Tumour location			1.000		1.000
Oesophagus	464 (86.4%)	239 (88.5%)		68 (87.2%)	
Stomach	73 (13.6%)	31 (11.5%)		10 (12.8%)	
cT stage			1.000		0.006
cT_1	9 (1.7%)	11 (4.1%)		7 (9.0%)	
cT_2	182 (33.9%)	87 (32.2%)		18 (23.1%)	
cT_3	297 (55.3%)	146 (54.1%)		46 (59.0%)	
cT_4	9 (1.7%)	3 (1.1%)		3 (3.8%)	
cT_X	40 (7.4%)	23 (8.5%)		4 (5.1%)	
cN stage			1.000		1.000
cN_0	264 (49.2%)	118 (43.7%)		33 (42.3%)	
cN_1	180 (33.5%)	98 (36.3%)		29 (37.2%)	
cN_2	86 (16.0%)	45 (16.7%)		13 (16.7%)	
cN_3	7 (1.3%)	9 (3.3%)		3 (3.8%)	
Treatment			1.000		<0.001
Chemoradiation (high dose)	48 (8.9%)	17 (6.3%)		5 (6.4%)	
Chemoradiation (low dose)	35 (6.5%)	21 (7.8%)		16 (20.5%)	
nCRT and surgery	324 (60.3%)	175 (64.8%)		34 (43.6%)	
nCT and surgery	37 (6.9%)	17 (6.3%)		6 (7.7%)	
Perioperative chemotherapy	16 (3.0%)	8 (3.0%)		5 (6.4%)	
Resection (primary tumour)	49 (9.1%)	19 (7.0%)		3 (3.8%)	
Other	28 (5.2%)	13 (4.8%)		9 (11.5%)	
Survival HR (95% CI)	1	0.92 (0.66–1.27)	0.601	5.34 (3.74–7.62)	<0.001

Supplementary Table 2: Exclusion analysis for the three-months analysis.
This analysis compares characteristics of included patients, patients who did not fill out the first questionnaire within 14 days after the start of treatment (late start), and patients who filled out their last questionnaire 14 days before the end of this period or earlier (early finish). nCRT: neoadjuvant chemoradiation; nCT: neoadjuvant chemotherapy.

	Included (N = 480)	Late start (N = 270)	p-value	Lost to follow-up (N = 135)	p-value
Age			0.822		0.639
Mean (SD)	66.0 (8.4)	65.0 (8.7)		67.2 (8.9)	
Sex			1.000		0.156
Female	100 (20.8%)	50 (18.5%)		40 (29.6%)	
Male	380 (79.2%)	220 (81.5%)		95 (70.4%)	
BMI			1.000		0.962
Mean (SD)	26.0 (4.4)	26.3 (4.2)		26.1 (3.9)	
WHO performance status			1.000		0.962
0	248 (62.0%)	136 (55.7%)		65 (61.3%)	
1	136 (34.0%)	96 (39.3%)		34 (32.1%)	
2	14 (3.5%)	10 (4.1%)		5 (4.7%)	
3+	2 (0.5%)	2 (0.8%)		2 (1.9%)	
Tumour location			1.000		0.687
Oesophagus	411 (85.6%)	239 (88.5%)		121 (89.6%)	
Stomach	69 (14.4%)	31 (11.5%)		14 (10.4%)	
cT stage			1.000		0.023
cT_1	8 (1.7%)	11 (4.1%)		8 (5.9%)	
cT_2	166 (34.6%)	87 (32.2%)		34 (25.2%)	
cT_3	266 (55.4%)	146 (54.1%)		77 (57.0%)	
cT_4	6 (1.2%)	3 (1.1%)		6 (4.4%)	
cT_X	34 (7.1%)	23 (8.5%)		10 (7.4%)	
cN stage			0.838		0.137
cN_0	244 (50.8%)	118 (43.7%)		53 (39.3%)	
cN_1	162 (33.8%)	98 (36.3%)		47 (34.8%)	
cN_2	67 (14.0%)	45 (16.7%)		32 (23.7%)	
cN_3	7 (1.5%)	9 (3.3%)		3 (2.2%)	
Treatment			1.000		<0.001
Chemoradiation (high dose)	41 (8.5%)	17 (6.3%)		12 (8.9%)	
Chemoradiation (low dose)	27 (5.6%)	21 (7.8%)		24 (17.8%)	
nCRT and surgery	296 (61.7%)	175 (64.8%)		62 (45.9%)	
nCT and surgery	31 (6.5%)	17 (6.3%)		12 (8.9%)	
Perioperative chemotherapy	9 (1.9%)	8 (3.0%)		12 (8.9%)	
Resection (primary tumour)	49 (10.2%)	19 (7.0%)		3 (2.2%)	
Other	27 (5.6%)	13 (4.8%)		10 (7.4%)	
Survival HR (95% CI)	1	1.20 (0.85–1.71)	0.307	6.57 (4.77–9.05)	<0.001

Supplementary Table 3: Exclusion analysis for the six-months analysis.

This analysis compares characteristics of included patients, patients who did not fill out the first questionnaire within 14 days after the start of treatment (late start), and patients who filled out their last questionnaire 14 days before the end of this period or earlier (early finish). nCRT: neoadjuvant chemoradiation; nCT: neoadjuvant chemotherapy.

	Three months			Six months		
	Correlation coefficient (ρ)	**Mean summary score at end of period (SD)**	**p-value**	**Correlation coefficient (ρ)**	**Mean summary score at end of period (SD)**	**p-value**
Sex			1.000			0.416
Female		77.64 (16.43)			75.60 (15.93)	
Male		79.72 (15.28)			78.24 (13.98)	
Age	-0.012		1.000	0.115		0.069
BMI	0.041		1.000	0.026		1.000
WHO performance status			0.138			0.017
0		80.59 (15.64)			78.89 (13.94)	
1		78.04 (14.73)			75.69 (14.20)	
2		70.15 (13.36)			67.16 (15.62)	
3+		69.45 (7.10)			61.84 (15.36)	
Education level			1.000			1.000
Higher education		80.35 (16.37)			77.65 (15.01)	
Lower vocational education		79.25 (14.58)			77.90 (13.86)	
Primary school education		77.74 (16.74)			76.85 (14.96)	
Secondary vocational education		79.00 (15.37)			77.98 (14.33)	
Tumour location			1.000			0.270
Oesophagus		78.97 (15.76)			77.17 (14.47)	
Stomach		81.04 (14.16)			80.79 (13.91)	
Treatment			1.000			0.786
Chemoradiation (high dose)		76.97 (14.84)			80.02 (16.14)	
Chemoradiation (low dose)		79.38 (18.28)			82.13 (17.50)	
nCRT and surgery		79.98 (15.45)			76.76 (13.76)	
nCT and surgery		78.95 (11.66)			75.41 (11.79)	
Perioperative chemotherapy		72.28 (16.74)			75.03 (18.87)	
Resection (primary tumour)		80.79 (15.32)			79.20 (14.40)	
Other		76.31 (18.17)			80.74 (16.32)	
Baseline summary score	0.583		<0.001	0.546		<0.001

Supplementary Table 4: Associations between predictor variables and summary score at three and at six months.
SD: standard deviation; nCRT: neoadjuvant chemoradiation; nCT: neoadjuvant chemotherapy.

Chapter 10

Informing patients with oesophagogastric cancer about treatment outcomes by using a web-based tool and training

A development and evaluation study

Loïs F. van de Water*, Héctor G. van den Boorn*, Florian Hoxha, Inge Henselmans, Mart M. Calff, Mirjam A.G. Sprangers, Ameen Abu-Hanna, Ellen M.A. Smets and Hanneke W.M. van Laarhoven

Based on:

L.F. van de Water*, H.G. van den Boorn*, F. Hoxha, *et al.* Informing patients with esophagogastric cancer about treatment outcomes by using a web-based tool and training: development and evaluation study. Journal of Medical Internet Research 23 (2021).

* *These authors contributed equally*

Abstract

Background

Due to the increasing use of shared decision-making, patients with oesophagogastric cancer play an increasingly important role in the decision-making process. To be able to make well-informed decisions, patients need to be adequately informed about treatment options and their outcomes, namely survival, side effects or complications, and health-related quality of life. Web-based tools and training programs can aid physicians in this complex task. However, to date, none of these instruments are available for use in informing patients with oesophagogastric cancer about treatment outcomes.
This study aims to develop and evaluate the feasibility of using a web-based prediction tool and supporting communication skills training to improve how physicians inform patients with oesophagogastric cancer about treatment outcomes. By improving the provision of treatment outcome information, we aim to stimulate the use of information that is evidence-based, precise, and personalised to patient and tumour characteristics and is communicated in a way that is tailored to individual information needs.

Methods

We designed a web-based, physician-assisted prediction tool, SOURCE, to be used during consultations by using an iterative, user-centred approach. The accompanying communication skills training was developed based on specific learning objectives, literature, and expert opinions. The SOURCE tool was tested in several rounds: a face-to-face focus group with six patients and survivors, semi-structured interviews with five patients, think-aloud sessions with three medical oncologists, and interviews with six field experts. In a final pilot study, the SOURCE tool and training were tested as a combined intervention by five medical oncology fellows and three oesophagogastric outpatients.

Results

The SOURCE tool contains personalised prediction models and data from meta-analyses regarding survival, treatment side effects and complications, and health-related quality of life. The treatment outcomes were visualised in a patient-friendly manner by using pictographs, barcharts and line graphs. The communication skills training consisted of blended learning for clinicians comprising an e-learning and two face-to-face sessions. Adjustments to improve both training and the SOURCE tool were made according to feedback from all testing rounds.

Discussion

The SOURCE tool and training could play an important role in informing patients with oesophagogastric cancer about treatment outcomes in an evidence-based, precise, personalised, and tailored manner. The preliminary evaluation results are promising and provide valuable input for the further development and testing of both elements. However, the remaining uncertainty about treatment outcomes in patients and established habits in doctors, in addition to the varying trust in the prediction models, might influence the effectiveness of the tool and training in daily practice. We are currently conducting a multi-centre clinical trial to investigate the impact that the combined tool and training have on the provision of information in the context of treatment decision-making.

Introduction

Oesophageal and gastric cancers rank eighth and fifth, respectively, in incidence worldwide.[1] The mortality rate is high and, even in the curative setting, the five-year survival rates do not exceed 50%.[2,3] Over the years, several treatment regimens have come into use, resulting in an array of treatments varying in their effectiveness regarding survival, health-related quality of life (HRQoL), and side effects and complications. For example, localised oesophageal cancer can be treated with resection, with or without neoadjuvant chemotherapy or chemoradiotherapy, or with definitive chemoradiation, and localised gastric cancer can be treated with resection with or without adjuvant or neoadjuvant chemotherapy.[4–6] Various options exist for metastasised cancers, with chemotherapy yielding the best survival rates. However, palliative radiotherapy and best supportive care may also be valuable options for specific groups of patients.[4,5,7–10]

Oftentimes, the choice between treatment options is based on preferences; the personal weighing of the pros and cons of the options plays a decisive role in the final decision made, and therefore, shared decision-making is needed.[11,12] For shared decision-making to be effective, patients need to be well-informed and thus be offered evidence-based and precise information on treatment outcomes. Evidence-based information refers to the best available, most accurate, and up-to-date evidence. Precise information is concrete, clear, and substantially detailed, such as "in five years, 45 out of 100 patients like you that are given this treatment will still be alive.". However, treatment outcomes can differ according to specific patient characteristics (such as age and performance status) and tumour characteristics (such as TNM staging and the number of metastases).[13,14] Thus, physicians face the challenge of having to inform patients on treatment-related outcomes in a manner that is not only evidence-based and precise but also personalised to the individual patient.

Physicians may face many other challenges when informing patients with cancer on treatment and related outcomes. A vast amount of information on the possible treatment options, including their procedures and associated risks and benefits, must be communicated within the time restrictions of a consultation.[15] Moreover, this information, including schedules, numbers, and probabilities, is often complex and therefore difficult for patients to process.[16,17] Patients' emotions can complicate information processing even further, especially as oesophagogastric cancer is a life-threatening disease.[18,19] Physicians consider dealing with these emotions as a difficult-to-acquire skill.[20] They of-

ten worry that their information might even increase a patient's anxiety or take away a patient's hope.[21–25] Therefore, physicians may have conflicting opinions and doubts about how to provide precise and numerical information regarding treatment risks and benefits.

Furthermore, tailoring the type and amount of information to the individual patient's information needs, interests, and concerns (e.g., one patient wants to be informed using exact percentages, whereas another would rather get a general description) has also been shown to be a difficult skill for physicians.[26] These challenges impede the ability to meet the information needs of patients with cancer.[27–29] Physicians rarely use clinical outcome data to systematically inform patients, given a certain treatment, on their chances of survival, the most likely side effects, and the consequences on their health-related quality of life.[30,31] However, it has been established that many patients want to receive more information on their treatment-related outcomes and want this information to be more precise.[32–36]

Several tools have been developed to aid physicians in this task by using prediction models to generate clarifying visualisations of personalised outcome data, such as the *PREDICT* and *Adjuvant! Online* tools for breast cancer.[37,38] To achieve personalised prediction, these models use multiple characteristics of the patient and the disease to create bar plots and Kaplan-Meier curves displaying survival data. However, to date, no web-based prediction tool exists for use in clinical consultations targeted at patients with oesophageal and gastric cancer.[39] Moreover, the probabilities of side effects and HRQoL related to the treatment options are not addressed in the current tools, although patients express information needs related to these outcomes.[32,40] Furthermore, several training programs are available to improve the communication skills of cancer care providers.[41,42] However, these often do not specifically address how to inform patients about treatment options and their particular outcomes, for instance, by using a prediction tool. Combining a prediction tool with communication skills training to address knowledge, attitudes, and skills might increase the usage and adoption of the new tool in clinical practice, improve the overall communication of outcome information by physicians, and stimulate shared decision-making.

Therefore, we aim to develop a web-based prediction tool and supporting communication skills training to improve how physicians inform patients with oesophagogastric cancer on treatment outcomes, namely, survival, side effects or complications, and HRQoL. To improve the provision of treatment outcome information, we aim for information that is evidence-based, precise, and personalised to the patient and tumour characteristics, and that is communicated in a way tailored to the individual information needs.

Introducing a change in physician-patient communication by adding a new instrument might initially result in resistance from users, as suggested by behaviour change theories.[43] For example, physicians might be reluctant to use the tool because it does not fit into their consultation routine or because they might lack trust in the prediction models. Therefore, our secondary aim is to evaluate the feasibility of the tool and training in practice by consulting physicians, patients, survivors, and experts and to iteratively improve the tool and training.

Methods

Both the tool and training were targeted at physicians in oncology who regularly conduct treatment decision-making consultations. In the development of the tool and training, we focused on patients with metastatic oesophageal and gastric cancer. With regard to shared decision-making, this group is confronted with the most complex decision-making process, where personal values and preferences play a large role in deciding among multiple relevant treatment options.

The iterative development and testing of this two-part intervention occurred in several phases following the 2008 Medical Research Council framework[44]. This framework provides guidance for developing complex interventions and presents several steps and elements necessary for the successful implementation of the intervention. The framework is divided into the following four phases: (1) development, (2) piloting, (3) evaluation, and (4) implementation. This study describes the first two phases: development and piloting. The development phase is described separately for the tool and training. Both elements of the intervention are joined in the piloting phase as a combined pilot study (see Figure 1 for an overview).

The web-based prediction tool: SOURCE

The web-based, physician-assisted prediction tool named SOURCE, which contains visualisations of evidence-based, precise, and personalised outcome information, was developed using an iterative, user-centred approach. The tool was designed to be used by oncology health care providers for decision-making consultations. The SOURCE tool, unlike other prediction tools such as *PREDICT* and *Adjuvant! Online*, was not designed for unsupervised use by patients at home to prevent incorrect use, misunderstanding, and lack of emotional support.

First, prediction models were developed to ensure that the SOURCE tool's treatment outcome information (survival, side effects and complications, and HRQoL) was evidence-based, precise, and personalised to the individual patient and tumour characteristics. Personalised predictions for survival using this tool are based on the SOURCE prediction models[45,46], which predict survival based on individual patient and tumour characteristics and are regularly updated when new data become available.

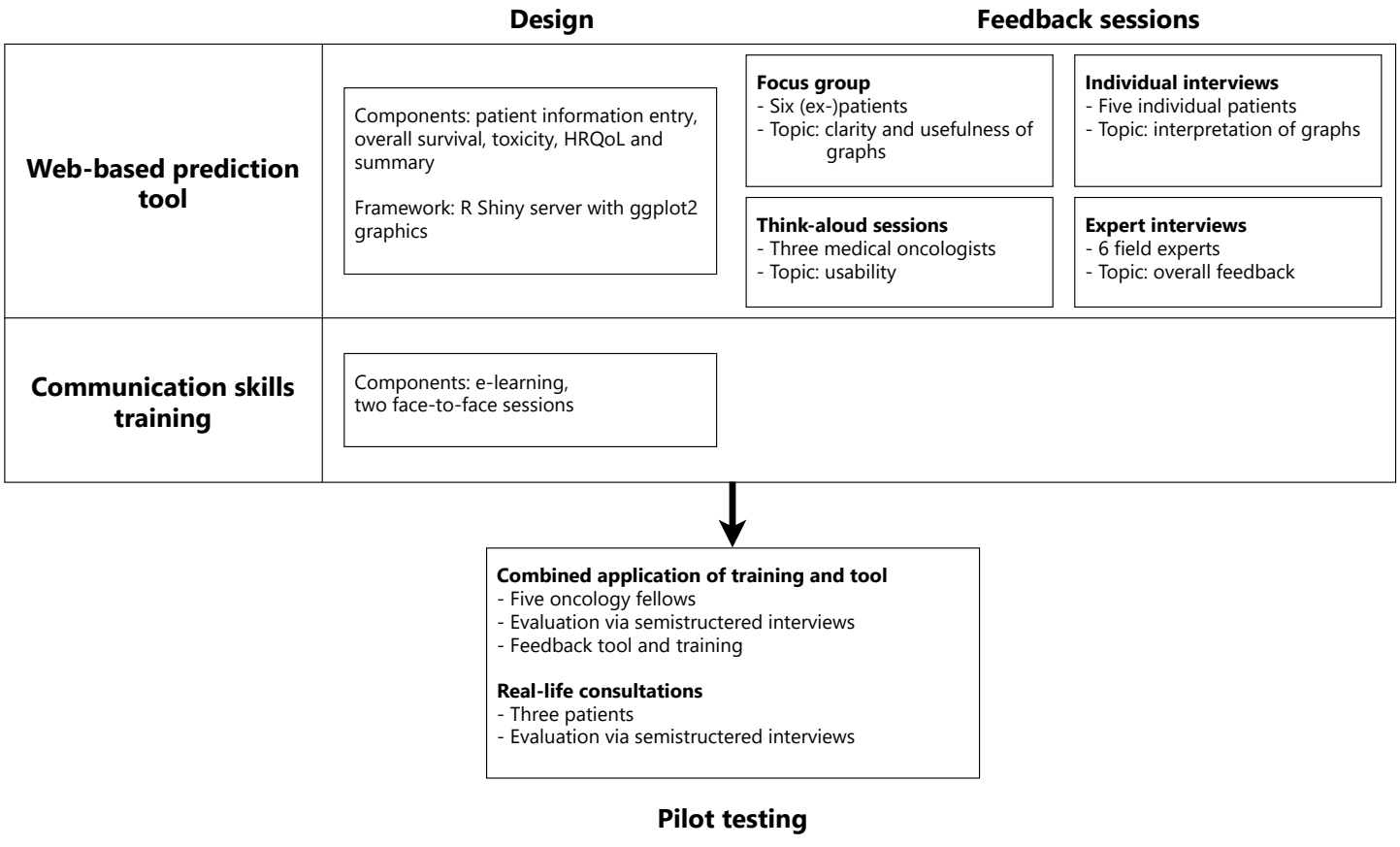


Figure 1: Schematic representation of the development process of the tool and training.
HRQoL: health-related quality of life.

Depending on the tumour location, either nine variables (for gastric cancer) or 13 variables (for oesophageal cancer) are required for predictions. These variables include patient characteristics, tumour staging, and metastasis characteristics. The model output is the probability of survival up to two years following diagnosis and allows for the comparison of multiple treatments.
The side effects (toxicity) of chemotherapy treatment are based on the TOXView meta-analyses.[7] These models establish the probability of adverse events such as nausea, alopecia, and neuropathy, stratified by mild or severe grade toxicity (according to the CTCAE[47]), for various chemotherapy regimens. These probabilities are not personalised to the individual characteristics and do not vary over time, as this was not possible with the available data.
Finally, predictions of HRQoL are available from meta-analyses and describe the change in the EORTC QLQ-C30 on the global health scale for best supportive care and chemotherapy in metastatic patients up to six months after diagnosis.[10]

Next, these models and meta-analyses were used to visualise treatment outcomes. For the visualisations to be easy to understand for patients, a previous systematic literature review on visual risk communication was consulted.[48] Furthermore, the literature about usability and usability guidelines for web-based applications[49–52] and existing prediction tools, such as *PREDICT* and *Adjuvant! Online*, were consulted[37,38]. On the basis of the literature, the first set of requirements for the tool was created according to the MoSCoW (Must have, Should have, Could have, Won't have) system.[53] This process resulted in the requirements listed in Table 1.

A prototype of the web-interface was created based on the literature guidelines and first requirements. The web-based tool was developed using the *shiny* package (version 1.2.0; RStudio) supplemented with *ggplot2* (version 3.2.1) to create visualisations.[54,55] The creation, evaluation, and improvement of the tool followed an iterative user-centred design framework, where feedback was gathered from end-users (patients and physicians) and experts. By iteratively updating the tool, we aim to provide improvements for the tool after each feedback session and avoid receiving the same feedback after each feedback round. A total of four feedback sessions were conducted from January 2018 to July 2018.

Must have	
1.	After opening the tool, a data entry form is shown to enter the variables needed for the prediction models.
2.	The data entry form is dynamic and shows only relevant variables.
3.	Survival, adverse events, and health-related quality of life outcomes are displayed in their own tabs, and only one outcome is displayed at a time.
4.	The outcomes are displayed graphically in a screen-filling image.
Should have	
1.	The data entry form contains input validation to avoid mistakes during entry.
2.	The data entry contains explanations of the variables.
3.	The plots can be tailored to the patients' and physicians' preferences (e.g., time frame and treatments to be compared).
Could have	
1.	The tool's display language can be set to Dutch or English.
2.	A textual summary accompanying the plots can be generated and printed so the patient can review the information at a later time.
3.	A help function for physicians is incorporated.

Table 1: Overview of the requirements of the web-interface according to the MoSCoW (Must have, Should have, Could have, Won't have) system.[53]

First, a face-to-face focus group was conducted with six patients with oesophageal and gastric cancer and survivors from the Foundation for Patients with Cancer in the Digestive system after verbal informed consent was provided. The aim of this focus group was to obtain feedback on the tool in a group setting and promote discussions among the group members. One of the researchers acted as a moderator and presented the participants with each of the tool's graphs. Each displayed graph was accompanied by a short oral explanation, after which the participants were asked for their opinions. Feedback on the web-interface and suggestions for improvement supported by multiple participants were used to create an improved version of the tool. The focus group session was audio-recorded and analysed according to micro-interlocutor[56] analysis to systematically evaluate the participants' remarks.

In a second feedback round, 30-minute, semi-structured, face-to-face interviews were conducted with individual patients with oesophageal and gastric cancer. The main focus of these audio-recorded interviews was to determine whether the patient's interpretation of the revised graphs was adequate. By conducting interviews with individual patients rather than in a group, the aim was to review patient interpretations without the influence of other patients. The interviews were conducted using a piloted script. A total of five outpatient participants were recruited by an oncologist at the Amsterdam University Medical Centres. Following the participants' informed consent, two researchers presented the tool to patients by using fictitious predictions of treatment outcomes. Patients were asked to interpret the presented graphs

and describe their meaning to assess their understanding. Thereafter, the researcher provided the correct description of the graph, and the patients' subsequent feedback was gathered. Feedback was registered in a response matrix, including the frequency of different remarks, to establish which possible improvements could be implemented.

The third feedback round aimed to evaluate the usability of the tool when used by medical oncologists, and three medical oncologists at the Amsterdam University Medical Centres participated in individual face-to-face think-aloud sessions. After providing informed consent, they were asked to use the tool for two paper patient cases while stating out loud whatever came to mind. The cases described fictitious patients with oesophageal or gastric cancer, including some of their clinical characteristics. Several tasks and questions about specific outcomes (e.g., "What is the one-year survival probability with best supportive care?" and "Which treatment has the best health-related quality of life after six months?") were posed to guide the use of the tool by medical oncologists. At the beginning of the think-aloud session, a video explaining the think-aloud method[57] was presented and the participants were asked to complete a short practice exercise to ensure that they understood the think-aloud method before starting the task. After the think-aloud session, participants completed the System Usability Scale (SUS)[52] to measure the ease of use and overall likeability of the web-based tool. Both screen captures and audio recordings were registered during the think-aloud session. One of the researchers (FH) used both recordings to register whether the oncologists successfully completed the tasks, how many mouse clicks they used to complete a task, and which buttons they clicked on the web-interface. The median SUS scores were calculated to provide a quantitative indication of usability.

In the fourth and last round, feedback from experts was gathered by conducting semi-structured interviews with six researchers with expertise in patient-physician interaction, shared decision-making, risk communication, medical informatics, and clinical decision support software. The experts were presented with a walkthrough of the tool and its options. Interviews were recorded and summarised to determine which possible improvements were brought forward.

The SOURCE supportive communication skills training

The communication skills training was developed to educate physicians on informing patients with cancer in a treatment decision-making consultation using the SOURCE tool. Due to the complexity of the skills needed, it was important to specify clear learning goals. As stated in complex learning theory, when training complex skills, the desired learning outcomes must address the following domains: knowledge, attitudes, and skills.[58–60] The training aimed for physicians to be able to name the most important tips and tricks for adequately informing patients on treatment outcomes and communicating treatment risks and benefits (knowledge). Furthermore, the training aimed for physicians to have a positive outlook on using numbers to inform patients about treatment outcomes and their ability to inform patients in an evidence-based, precise, personalised, and tailored manner (attitude). Moreover, the training aimed for physicians to be able to use the SOURCE tool and to incorporate the tool to inform patients during consultations (skills). Finally, the training aimed to increase physicians' ability to provide information tailored to patients' informational needs and level of understanding (skills). A team (N = 5) of experts in medical communication and psycho-oncology and experienced trainers in medical communication discussed the context and content of the training and set learning objectives. In addition, the literature on training and shared decision-making frameworks was reviewed.

As physicians value time-efficient and flexible training, the training was designed as blended learning, encompassing preparatory e-learning and a face-to-face component. The four-step shared decision-making model proposed by Stiggelbout et al.[12] was used as a framework. This model distinguishes the following four essential steps for shared decision-making: (1) setting the agenda, (2) informing about treatment options, (3) exploring patients' values, and (4) making a decision in agreement. The outline of the training was based on a previous communication skills training for skills in shared decision-making, as designed for and proven to be effective in the CHOICE trial[61] and the literature on the guidelines for effective communication skills training.[42,62] The focus of the SOURCE training is the second step of this model, that is, informing patients about treatment options and the pros and cons thereof.

E-learning

First, the e-learning was targeted at summarising the evidence base for effective information provision and providing physicians with tips and tricks for clinical practice. To this end, we consulted the literature related to theories, evidence, and guidelines on the provision of information in medical practice.[12,63–67] The assembled literature and theories were summarised into short chapters, each covering a different subtopic. The expert team discussed the scripts in these chapters to obtain a consensus on the frameworks and models used. Interactive elements, such as exercises, were added to the e-learning to enable the learner to actively process the information. Second, the e-learning aimed to introduce the SOURCE tool, thereby addressing the use and functionalities of the tool and the underlying prediction models.

An earlier study concluded that physicians value both visual attractiveness and variation between learning activities in e-learnings[16]; therefore, the layout, animations, and videos were developed in cooperation with a small visual design company, Public Cinema.

Face-to-face sessions

Face-to-face sessions were developed based on previous experience in developing and evaluating communication skills training in oncology.[42,61,68] The most important recommendations from these earlier studies were to role-play with an actor to practice the lessons learned during the training and provide the trainee with personal feedback.[42] Development took place in multiple sessions with the expert team. The basic assumptions for effective information provision, as incorporated in the e-learning, served as a starting point for the training content. Derived ideas were written down and discussed to create a training script and a supportive PowerPoint presentation. The casuistry for the training actors was developed together with a clinical expert (HvL). These multiple development sessions led to a conceptual version of the training.

Pilot study tool and training

A pilot study was conducted from December 2018 to March 2019 to test both the tool and training in a real-life setting. As this pilot study targeted patients with advanced disease only, we included medical oncologists and metastatic cancer patients as study participants. The pilot study was evaluated by the Medical Ethics Review Committee of the Academic Medical Centre Amsterdam (W18278). In total, five medical oncology fellows (two men and three women) from two university medical centres were invited to use the tool and trained according to the concept training format. After completion of the training, participating fellows were individually interviewed via telephone

in a semi-structured manner to gather feedback to improve both the SOURCE tool and training. For the tool, the focus of the feedback was on opinions and experiences regarding usability and willingness to use the tool. Regarding the training, feedback on the different components, the training as a whole, and perceived utility were collected.

In addition, an experienced medical oncologist (HvL) conducted three treatment decision-making consultations with outpatients using the SOURCE tool for information provision and in line with the training principles. These consultations were recorded on video after obtaining written informed consent from the patients and oncologist. To comply with ethical standards and according to the training, only information that the patient wanted to receive was disclosed to the patient. One-on-one semi-structured interviews were conducted with the three patients by one of the researchers (LvdW) to gather their experiences with the physician's outcome information and the use of the SOURCE tool.

Results

The web-based prediction tool: SOURCE

A prototype web-interface was created based on the findings of a systematic review of the effects of different types of risk communication on patients with cancer.[48] Following this review, we decided to use clear and precise risk information (e.g., percentages or frequencies) and simple graphs with a limited amount of information displayed. As the review did not yield consistent guidelines on which types of graphs to use, it was decided to visualise the outcomes in multiple ways. In this way, graphs can be used according to the preferences of individual patients, and the amount and presentation format of the information displayed can be tailored to their needs and preferences. The resulting *RShiny* web-interface runs on an x64 Linux server (version 3.10.0).

SOURCE tool components

The final tool, SOURCE 1.103, contains five main components. The first component, the patient information entry component, allows the oncologist to enter the patient characteristics necessary for the prediction models and meta-analyses, using supporting information, such as the definitions of TNM variables (Figure 2).

The survival component was visualised in two ways: an icon array displaying the survival probability at a given point in time by colouring a subset of 100 figures and a Kaplan-Meier curve (line graph) displaying the survival probability over time (Figure 3). The survival component incorporates the possibility of switching between the two presentation formats. From the options menu, it is possible to select specific treatments for comparison (e.g., best supportive care and chemotherapy), to change the time frame of the prediction (from six to 24 months), to show a per treatment confidence interval and to visualise three survival scenarios per treatment (indicating the best-case scenario comprising the top 25%, the worst-case scenario comprising the bottom 25%, and the typical outcome comprising the middle 50%)[63,64].

The side effects component displays bar charts for various toxicities (Figure 4), as the meta-analysis provided static probabilities for each of the adverse events.[7] Each side effect was visualised by two stacked bars, one for mild side effects and one for severe side effects on the CTCAE scale.[69] The side effects of multiple chemotherapy regimens can also be compared. To avoid information overload, only the three most frequently occurring side effects are shown initially, although it is possible to display all side effects.

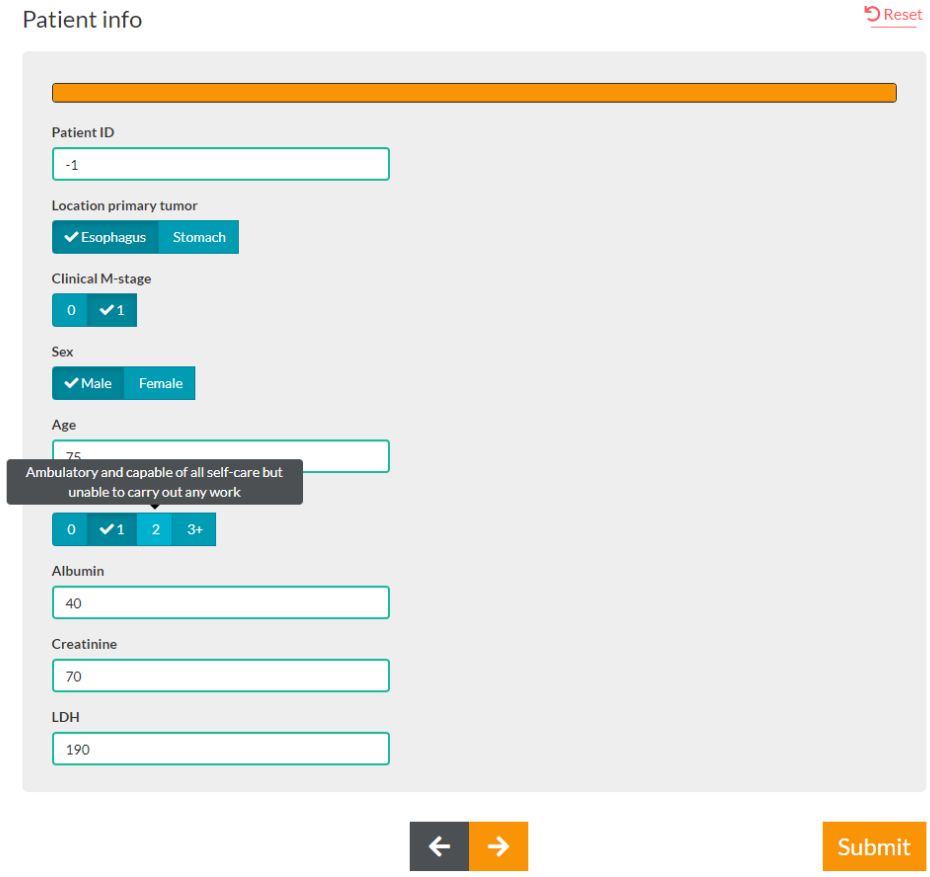


Figure 2: A screenshot of patient data entry.
The data entry screen displays the fields that are necessary for the prediction models and meta-analyses. Additional information on variables such as the WHO performance status is provided with a mouse-over.

HRQoL is displayed in a line graph and shows the EORTC QLQ-C30 global health score over time (Figure 5). There are options to compare HRQoL in best supportive care with chemotherapy and display a confidence interval and reference value (obtained from the EORTC reference values manual[70]).
The final component is a summary that can be printed as a handout for the patient or saved as a PDF file. This feature enables physicians to show the aforementioned graphs accompanied by an explanatory text. This text is dynamically generated using the selected treatment data and explains the content of the graphs.

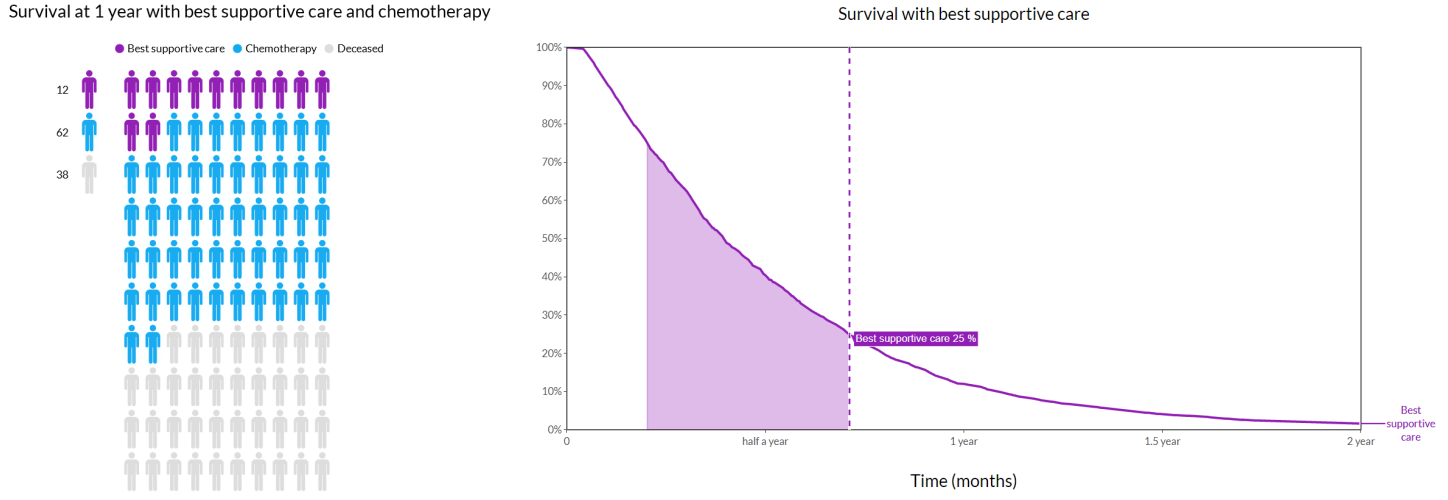


Figure 3: A screenshot of survival graphs.
On the left, a pictograph displaying the predicted survival for best supportive care and chemotherapy after 1 year is shown. On the right, the Kaplan-Meier curve for the best supportive care is shown. The optional shaded area displays the so-called typical outcome scenario (with survival ranging from 25% to 75%).

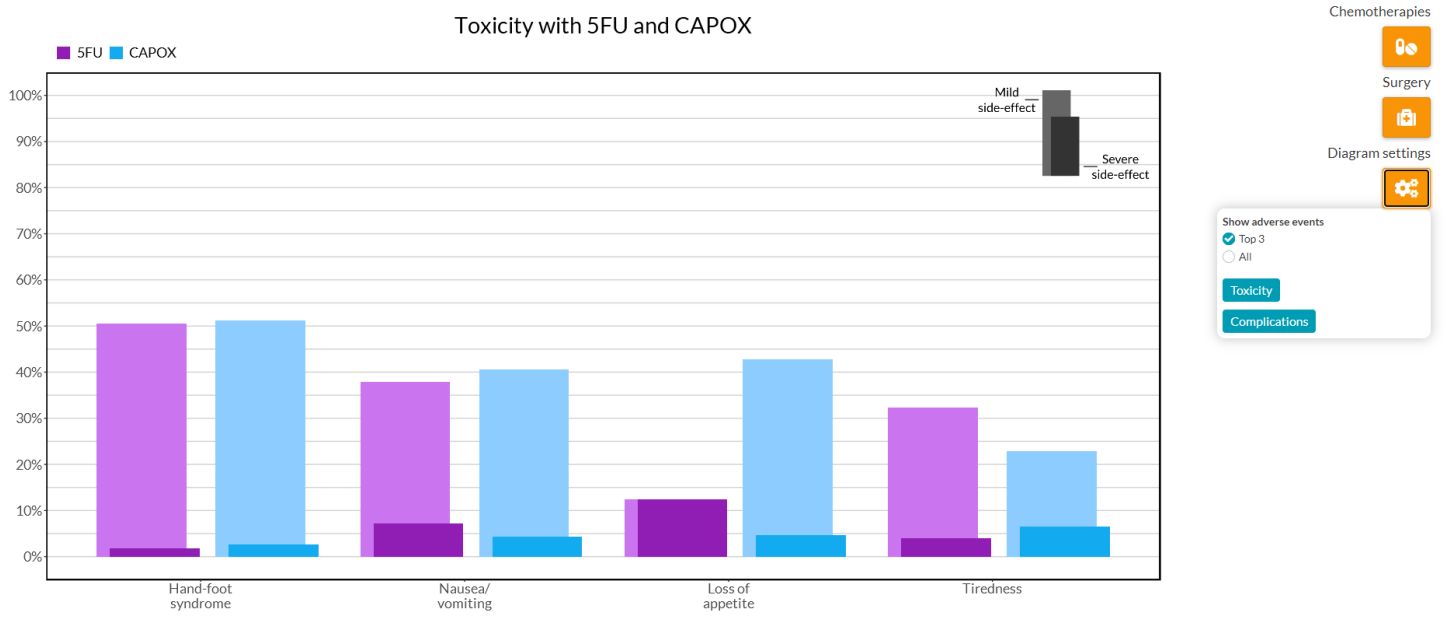


Figure 4: A screenshot of the side effects bar chart.
This displays the three most commonly occurring toxicities for both 5FU and CAPOX. The darker bars indicate severe toxicities, and the lighter bars indicate mild toxicities. 5FU: 5-fluorouracil; CAPOX: capecitabine combined with oxaliplatin.

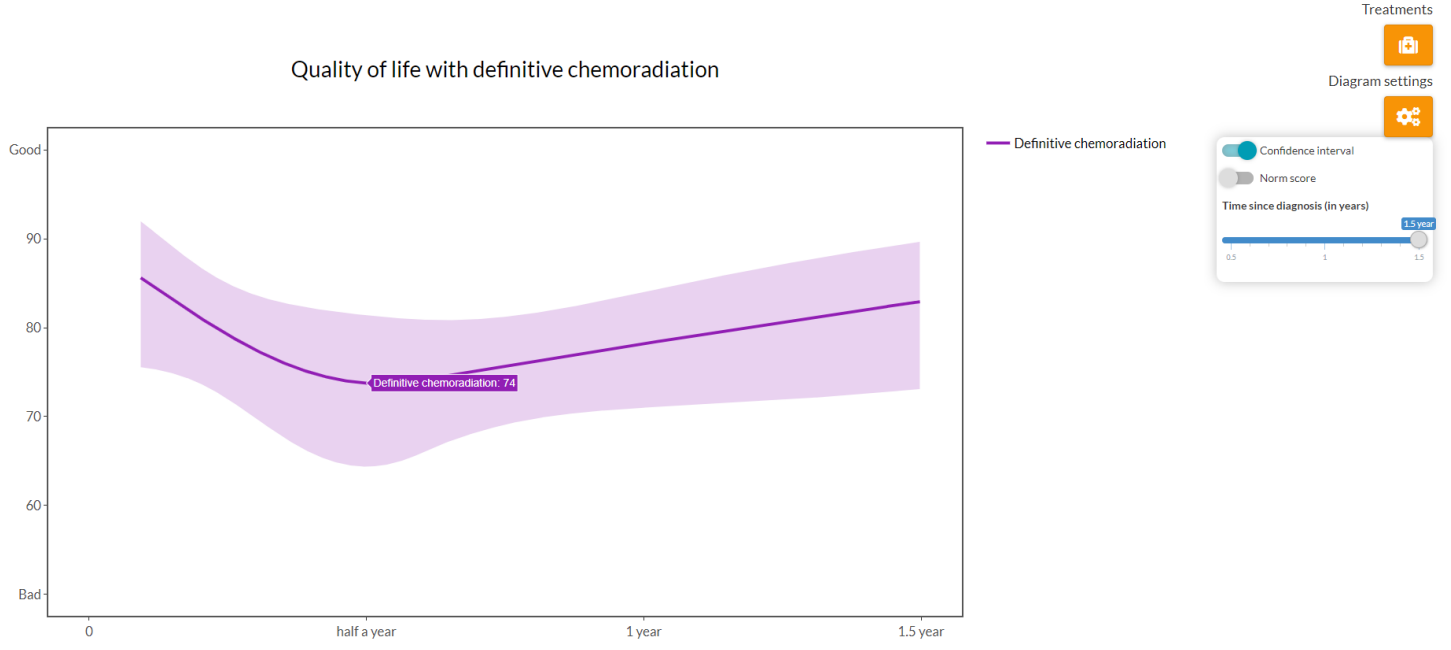


Figure 5: A screenshot of the HRQoL line graph.
The graph displays the HRQoL following definitive chemoradiation. The shaded area optionally displays the confidence interval of the HRQoL line.

Evaluation round feedback

The four evaluation rounds of the tool resulted in several minor visual and functional adjustments to the web-interface, as described in Supplementary Table 1. Major adjustments to the tool resulting from the gathered feedback were mostly adjustments regarding usability, such as increasing the font size and positioning of the legend. For the survival outcome, the icon array was found to be the most comprehensible, whereas the line graph provided the most insight for survival over time. Therefore, it was decided to keep both formats for the survival outcome, as the graphs supplemented each other. The line graph also remained in the tool as it incorporated the scenario's functionality (indicating worst-case, typical, and best-case survival), a feature that was found important by most patients. For side effects, it was found that the patients did not correctly interpret the meaning of the stacked bar charts. The bars were changed into two non-stacked bars with 90% overlap to display mild and severe adverse events for the same side effects to increase clarity on the meaning of the graph (Figure 4). Furthermore, it was decided to remove bar charts as a display option for HRQoL data, as both patients and oncologists found the graph unclear and wanted the data to be displayed over time, as HRQoL may increase and decrease over time. Showing the predictions at a single time point may therefore not provide sufficient insight. In the final design, various options are available to personalise the displayed graphs. Regarding the usability of the tool, it received a median SUS score of 90.0 out of three ratings (above the 'excellent' threshold of 80.3 points[52]) during the think-aloud sessions.

In the pilot study, four out of five oncology fellows participated in an evaluative semi-structured interview following the training. The fifth fellow did not respond to repeated invitations. All four oncology fellows reported that the tool was highly usable overall. Some minor suggestions were provided for improving the display of certain options, graphs, and buttons. Of the four oncologists, three reported that they would use the tool in their clinical practice. They especially valued the personalised nature of the tool's predictions and the clear and easy-to-understand visualisations for patients. Furthermore, the inclusion of HRQoL data and the option to print a summary for the patient were found useful. The fourth fellow would like the prediction models to be further developed before using the tool. For instance, during the pilot training and interviews, several critical remarks on the prediction models were expressed, such as the lack of WHO performance status as a predictor of overall survival. These comments likely reflect a possible lack of trust in the underlying models and analyses of the tool. Another barrier to using the tool that one

of the fellows addressed was the fear of emotionally confronting the patient with the exact numbers. This fellow did indicate that this fear was overcome through the tailoring skills that he had acquired during the training.

The SOURCE supportive communication skills training

The e-Learning module

A short and to-the-point e-learning module was developed to provide an overview of the theoretical background on the provision of information in a treatment decision context and introduce a web-based prediction tool. The e-learning module starts with a short peer endorsement video of the training of physicians that discusses the SOURCE tool. Subsequently, physicians can navigate through four chapters. The first chapter provides an overview of the principles of effective information provision in the context of a shared decision. The second chapter introduces the physicians to the SOURCE tool by presenting them with a tailor-made instructional video of its use and functions. Furthermore, a summary of the tool's models and their underlying data is provided. The third chapter provides an overview of tips and tricks for informing patients about the risks and benefits of treatment (Supplementary Table 2). The final chapter consists of a short and practical summary of key take-home messages. In all chapters, textual information and short assignments are alternated by instructional videos and animated knowledge clips. A simple and appealing visual design is applied.

Face-to-face training sessions

The face-to-face component of the training consisted of two group sessions of 3.5 hours each, provided by an experienced trainer, with approximately two to three weeks in between to facilitate intermediate practice. The sessions were aimed at small groups of two to six participants, as this approach enables every participant to practice and receive personal feedback. Such a setting can also promote interactivity.[71] Both sessions involved individual role-play exercises with a professional actor in which feedback was provided by the trainer, the actor, and peers to learn additional skills.[62] Furthermore, group discussions were stimulated and led by the trainers. This approach was used to encourage physicians to discuss their attitudes toward using numbers and the tool in the context of the provision of information to patients.[62,72]

The first session covered the skills of setting the agenda of the decision-making consultation, introducing the tool and informing patients on survival outcomes. Physicians were asked to practice separate parts of the consultation while receiving feedback from other physicians and the trainer. Physicians were instructed to use the acquired knowledge and skills during their outpatient consultations before the next training session. The second session allowed for the repetition of issues addressed during the first session and sharing experiences of applying the lessons learned of the first session in clinical practice. Next, skills were addressed, again with role-play and feedback, to inform patients about treatment outcomes in terms of side effects and complications and HRQoL. This session also addressed how to conclude decision-making consultations.

In both sessions, tailoring of the amount and type of outcome information to specific patients played a significant role in both role-play practice and group discussions. Tips and tricks were discussed regarding how to determine an individual's information needs and wants, how to fit these needs with the informational needs of the physician, and whether and how the tool could contribute to tailored information giving. The total duration of the blended learning was 7.5 study hours, which consisted of half an hour of e-learning and seven hours of face-to-face training.

Feedback and major adjustments

Fellows reported that they enjoyed participating in the training and specifically valued personal coaching and practical tips. Furthermore, fellows appreciated the trainers and actors. In their opinion, there was a good balance between the information provided by trainers and practical exercises. Fellows especially appreciated the feedback during the training from the actor and trainer on their role-play with the actor. Overall, the training was described as useful, and specific improvements were suggested.

A point of improvement that was brought up was the substantial time investment in the training. In particular, the pace of the e-learning and instructional video was considered too slow. Furthermore, the timing of the face-to-face sessions following a day of work was considered inconvenient. The most important adjustments to the training as a result of the fellows' feedback were related to accelerating the pace and adding an individual booster session to the training in which the physician could receive personal feedback from one of the trainers on a full, recorded consultation.

Despite their emotions on the subject, two out of the three patients were willing to participate in a short interview about their experiences with the consultation. Both appreciated the use of the SOURCE tool and the physicians' explanations of possible treatment outcomes. Patients expressed differences in their experiences regarding the amount of information about treatment options and outcomes. Although one patient reported being satisfied with the amount of information, the other indicated that the amount of information was too extensive for him to memorise it all. He needed a printed summary of the tool for support. Both patients mentioned their struggle with the meaning of risk or benefit for themselves, as great uncertainty remains about their own future, despite the information provided.

Discussion

Our study shows the iterative development and pilot testing of the SOURCE tool and training. This combined intervention was developed using scientific evidence and input from physicians, patients, and experts. This process resulted in the first web-based prediction tool to inform patients with oesophagogastric cancer during consultations on survival, side effects, and HRQoL of different treatment options. Furthermore, we created a supporting training to teach physicians the communication skills needed to use the tool and to provide patients with information in an evidence-based, precise, personalised, and tailored manner. Preliminary evaluation results are promising and provide valuable input for further development and testing of both elements.

Both the tool and training were valued by participating physicians and patients. Physicians especially appreciated the practical approach of the training; the multiple practice opportunities and personal feedback helped them use the tool. Nevertheless, despite their positive attitudes toward the tool and training, old habits could stand in the way of using the tool and may impede the use of learned communication techniques in clinical practice. Behavioural change theories show that many factors can contribute to, but also stand in the way of learning new behaviours. Resistance could, for instance, arise as a result of a different expected outcome of the tool or because of a low tolerance for change.[43] The transfer of training describes the possible behavioural change resulting of an educational intervention such as training. From the literature, we know that although certain trainee characteristics (such as the perceived utility of the training) and training design factors (such as a realistic training environment) can promote the transfer of training, they are also strongly influenced by characteristics of the work environment, such as situational cues (e.g., social support from peers or supervisors) or consequences (e.g., negative or strong emotional reactions from patients).[73] These characteristics can be difficult to control in the setting of everyday hospital care. However, concerning the training design factors, the distribution of the training sessions might be an important factor contributing to the transfer of this tool and training for daily clinical practice. Indeed, the so-called spacing between the two face-to-face sessions and the booster session of the training might help increase task performance.[74–79]

During the pilot training, it was noted that in some cases, the fellows lacked trust in the prediction models used in the tool. Further steps were taken to increase the physicians' trust. For instance, details about the underlying data and publications were added to the tool to provide more information about the methodology and sources. Model updates (such as the 2020 version of the survival model), which increase the model's performance and sample size and include, for instance, WHO performance status as a predictor, may also increase trust.[39] Finally, external validation of the models can also generate trust in the validity and applicability of the tool.[80]

The use of the SOURCE tool could also be influenced by the application's usability and how well the tool solves the patient-informing problem as perceived by patients and physicians. Therefore, iterative usability testing is necessary to achieve an acceptable level of usability. As the number of testing rounds in this study is limited compared to other studies[37,38], the tool's usability may be further improved. This issue will be an ongoing point of attention that we will address during future testing and development of the tool. From a patient's perspective, some uncertainties regarding treatment outcomes may be reduced during the consultation, whereas other uncertainties remain. For example, the tool might support patients in participating in shared decision-making, but active participation in this difficult choice might also overwhelm them.[81] These issues can be addressed in medical education and training by dealing with a broad spectrum of patient uncertainty.

On the basis of our experiences, we can provide several recommendations to aid future research in creating and evaluating web-based prediction tools with training. First, we advise involving end-users, such as patients and physicians, in the early stages of development. Assumptions and implementations are often made from the perspective of developers, which may not coincide with the needs and wishes of patients or physicians. By evaluating at an early stage, it is possible to adjust the tool and training, and subsequent improvements can be implemented more seamlessly. Although not formally evaluated in this study, user research to investigate patients' and physicians' ideas and expectations regarding such a tool could also contribute to the usability and adoption of the tool in clinical practice. Second, evaluation by physicians may be complicated because of their busy schedules. We recommend making the feedback rounds with physicians as short as possible, planning them sufficiently in advance, and having them take place on training days. Third, as insights on data visualisation and risk communication may change constantly, we recommend facilitating ongoing updates of a developed prediction tool. We also suggest

that future research should use the current state of the art when designing a new tool or training. Fourth, we advise that communication on outcomes with subjective interpretations, such as HRQoL, deserves a more prominent place in communication skills training. We noticed in our training that physicians often had trouble explaining outcomes such as HRQoL. Finally, it was observed that end-users sometimes had conflicting opinions regarding improvements to the tool or training. As it is not possible to cater to everyone's wishes, we recommend weighing the pros and cons of suggestions and deciding whether a personalisation option will be implemented (such as displaying survival as a pictograph and a line graph) or whether a single option will be implemented (such as the background colour of the web-interface).

Most of the patients and physicians who participated in this study agreed that the tool and training added value to clinical practice. However, bias may have played a role in the evaluation of the tool and training, as the evaluation only partly took place in clinical practice. To investigate this potential bias and the extent to which the combined tool and training aid in information provision in the context of treatment decision-making, we are currently performing the third phase of the Medical Research Council framework (evaluation) with a multi-centre effect study (registered under NCT04232735, the SOURCE trial). In this stepped-wedge trial, physicians receive training and use the SOURCE tool both in simulated patient assessments and with outpatients. The effect that the intervention has on the outcome information provided by oncology physicians is quantitatively investigated by recording these consultations before and after the intervention and analysing physicians' outcome-related remarks. The primary outcome of this study is the provision of precise outcome information; secondary outcomes include the amount of tailoring to the information needs of patients, the patients' own knowledge and opinions on the communicated outcome information, and the influence that the consultation has on patients' emotions. As trust in the SOURCE models was found to be a potential barrier to using the tool in the pilot study, physicians' trust in the models will be closely monitored and specifically addressed during the trial. The models included in the SOURCE tool are being continuously improved and updated, in part, to address these issues.

In this trial, both palliative and curative patients will participate, and models aimed at potentially curable patients (cM_0) will be added to the SOURCE tool. The survival models are based on the 2020 version of the models, which include updated palliative prediction models and newly developed curative prediction models for both oesophageal and gastric cancer.[45] The HRQoL model for curatively treated patients originates from a systematic review and meta-analysis[82], and treatment side effects for curatively treated patients were provided by the COMplot study[83]. The addition of these models to the SOURCE tool enables evidence-based and precise information personalised to the individual's characteristics in the full spectrum of patients with oesophagogastric cancer. As the tool is currently being tested in a trial, access is currently restricted to trial participants. However, after the conclusion of the trial, the tool will be freely available in both Dutch and English, enabling the use of the SOURCE tool in clinical practice.[84]

To conclude, we developed and evaluated a web-based tool and training to inform patients with oesophageal or gastric cancer regarding treatment outcomes. Through evaluation and a pilot study, patients and physicians indicated the added value of the tool and training, and both were improved based on their feedback. The tool and training are currently being evaluated in a multi-centre trial to determine their added value in clinical practice.

References

1. Ferlay, J., Soerjomataram, I. & Dikshit, R. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 136, E359–386 (2015).
2. Badgwell, B., Das, P. & Ajani, J. Treatment of localized gastric and gastroesophageal adenocarcinoma: The role of accurate staging and preoperative therapy. Journal of Hematology and Oncology 10, 1–7 (2017).
3. Van Hagen, P., Hulshof, M., van Lanschot, J., *et al.* Preoperative Chemoradiotherapy for Esophageal or Junctional Cancer. New England Journal of Medicine 366, 2074–2084 (2012).
4. Smyth, E. C., Verheij, M., Allum, W., *et al.* Gastric cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Annals of Oncology 27, v38–v49 (2016).
5. Lordick, F., Mariette, C., Haustermans, K., *et al.* Oesophageal cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Annals of Oncology 27, v50–v57 (2016).
6. Van den Ende, T., Ter Veer, E., Machiels, M., *et al.* The efficacy and safety of (neo) adjuvant therapy for gastric cancer: a network meta-analysis. Cancers 11, 80 (2019).
7. Ngai, L. L., ter Veer, E., van den Boorn, H. G., *et al.* TOXview: a novel graphical presentation of cancer treatment toxicity profiles. Acta Oncologica 58, 1138–1148 (2019).
8. Ter Veer, E., Mohammad, N. H., Van Valkenhoef, G., *et al.* The Efficacy and Safety of First-line Chemotherapy in Advanced Esophagogastric Cancer: A Network Meta-analysis. Journal of the National Cancer Institute 108 (2016).
9. Ter Veer, E., Mohammad, N. H., van Valkenhoef, G., *et al.* Second-and third-line systemic therapy in patients with advanced esophagogastric cancer: a systematic review of the literature. Cancer and Metastasis Reviews 35, 439–456 (2016).
10. Van Kleef, J. J., Ter Veer, E., Van den Boorn, H. G., *et al.* Quality of Life during Palliative Systemic Therapy for Esophagogastric Cancer: Systematic Review and Meta-Analysis. Journal of the National Cancer Institute 112, 12–29 (2020).
11. O'Connor, A. M., Légaré, F. & Stacey, D. Risk communication in practice: the contribution of decision aids. The BMJ 327, 736–740 (2003).
12. Stiggelbout, A. M., Pieterse, A. H. & De Haes, J. C. Shared decision making: Concepts, evidence, and practice. Patient Education and Counseling 98, 1172–1179 (2015).
13. Van den Ende, T., Ter Veer, E., Mali, R., *et al.* Prognostic and predictive factors for the curative treatment of esophageal and gastric cancer in randomized controlled trials: a systematic review and meta-analysis. Cancers 11, 530 (2019).
14. Ter Veer, E., van Kleef, J. J., Schokker, S., *et al.* Prognostic and predictive factors for overall survival in metastatic oesophagogastric cancer: A systematic review and meta-analysis. European Journal of Cancer 103, 214–226 (2018).
15. Dugdale, D. C., Epstein, R. & Pantilat, S. Z. Time and the patient–physician relationship. Journal of general internal medicine 14, S34 (1999).
16. Stuij, S. M., Labrie, N. H. M., van Dulmen, S., *et al.* Developing a digital communication training tool on information-provision in oncology: uncovering learning needs and training preferences. BMC medical education 18, 1–12 (2018).
17. Politi, M. C., Han, P. K. J. & Col, N. F. Communicating the uncertainty of harms and benefits of medical interventions. Medical Decision Making 27, 681–695 (2007).
18. Visser, L. N. C., Tollenaar, M. S., van Doornen, L. J. P., *et al.* Does silence speak louder than words? The impact of oncologists' emotion-oriented communication on analogue

patients' information recall and emotional stress. Patient education and counseling 102, 43–52 (2019).
19. Kessels, R. P. C. Patients' memory for medical information. Journal of the Royal Society of Medicine 96, 219–222 (2003).
20. Clayton, J. M., Adler, J. L., O'Callaghan, A., *et al.* Intensive communication skills teaching for specialist training in palliative medicine: development and evaluation of an experiential workshop. Journal of palliative medicine 15, 585–591 (2012).
21. Politi, M. C. & Légaré, F. Physicians' reactions to uncertainty in the context of shared decision making. Patient education and counseling 80, 155–157 (2010).
22. McIntosh, J. Processes of communication, information seeking and control associated with cancer: A selective review of the literature. Social Science & Medicine (1967) 8, 167–187 (1974).
23. Kodish, E. & Post, S. G. Oncology and hope. Journal of Clinical Oncology 13, 1817–1822 (1995).
24. Good, M.-J. D. The practice of biomedicine and the discourse on hope. 1991.
25. Daugherty, C. K. & Hlubocky, F. J. What are terminally ill cancer patients told about their expected deaths? A study of cancer physicians' self-reports of prognosis disclosure. Journal of Clinical Oncology 26, 5988 (2008).
26. Douma, K. F. L., Koning, C. C. E., De Haes, H. C. J. M., *et al.* Do radiation oncologists tailor information to patients needs? And, if so, does it affect patients? Acta Oncologica 51, 512–520 (2012).
27. Oerlemans, S., Husson, O., Mols, F., *et al.* Perceived information provision and satisfaction among lymphoma and multiple myeloma survivors—results from a Dutch population-based study. Annals of hematology 91, 1587–1595 (2012).
28. Hack, T. F., Degner, L. F. & Parker, P. A. The communication goals and needs of cancer patients: a review. Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer 14, 831–845 (2005).
29. Puts, M. T. E., Papoutsis, A., Springall, E. & Tourangeau, A. E. A systematic review of unmet needs of newly diagnosed older cancer patients undergoing active cancer treatment. Supportive Care in Cancer 20, 1377–1394 (2012).
30. Parameswaran, R., McNair, A., Avery, K. N. L., *et al.* The role of health-related quality of life outcomes in clinical decision making in surgery for esophageal cancer: a systematic review. Annals of surgical oncology 15, 2372–2379 (2008).
31. Koedoot, C. G., Oort, F. J., De Haan, R. J., *et al.* The content and amount of information given by medical oncologists when telling patients with advanced cancer what their treatment options are: Palliative chemotherapy and watchful-waiting. European Journal of Cancer 40, 225–235 (2004).
32. Henselmans, I., Jacobs, M., van Berge Henegouwen, M. I., *et al.* Postoperative information needs and communication barriers of esophageal cancer patients. Patient education and counseling 88, 138–146 (2012).
33. Henselmans, I., Smets, E. M. A., Han, P. K. J., *et al.* How long do I have? Observational study on communication about life expectancy with advanced cancer patients. Patient education and counseling 100, 1820–1827 (2017).
34. Zandbelt, L. C., Smets, E. M. A., Oort, F. J., *et al.* Satisfaction with the outpatient encounter. Journal of general internal medicine 19, 1088–1095 (2004).
35. Jacobs, M. Supporting Patients in obtaining and Oncologists in providing Evidence-Based Health-Related Quality of Life Information prior to and after Esophageal Cancer Surgery. PhD thesis (2015).

36. Hodgkinson, K., Butow, P., Hunt, G. E., *et al.* The development and evaluation of a measure to assess cancer survivors' unmet supportive care needs: TheCaSUN (Cancer Survivors' Unmet Needs measure). Psycho-Oncology 16, 796–804 (2007).
37. Ravdin, P. M., Siminoff, L. A., Davis, G. J., *et al.* Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. Journal of Clinical Oncology 19, 980–991 (2001).
38. Wishart, G. C., Azzato, E. M., Greenberg, D. C., *et al.* PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Research 12, 1–10 (2010).
39. Van den Boorn, H. G., Engelhardt, E. G., van Kleef, J., *et al.* Prediction models for patients with esophageal or gastric cancer: A systematic review and meta-analysis. PLoS ONE 13 (2018).
40. Farmer, G. D., Pearson, G. M., Skylark, W. J., *et al.* Making prognostic algorithms useful in shared decision-making: Patients and clinicians requirements for the Predict: Breast Cancer interface. medRxiv (2020).
41. Butler, L., Degner, L., Baile, W. & Landry, M. Developing communication competency in the context of cancer: a critical interpretive analysis of provider training programs. Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer 14, 861–872 (2005).
42. Bos–van den Hoek, D. W., Visser, L. N. C., Brown, R. F., *et al.* Communication skills training for healthcare professionals in oncology over the past decade: a systematic review of reviews. Current opinion in supportive and palliative care 13, 33–45 (2019).
43. Kotter, J. P. & Schlesinger, L. A. Choosing Strategies for Change. Harvard Business Review 86 (2008).
44. Craig, P., Dieppe, P., Macintyre, S., *et al.* Developing and evaluating complex interventions: the new Medical Research Council guidance. The BMJ 337 (2008).
45. Van den Boorn, H. G., Abu-Hanna, A., Haj Mohammad, N., *et al.* SOURCE: Prediction models for overall survival in patients with metastatic and potentially curable esophageal and gastric cancer. Journal of Clinical Oncology 38, 301–301 (2020).
46. Van den Boorn, H. G., Abu-Hanna, A., Ter Veer, E., *et al.* SOURCE: A Registry-Based Prediction Model for Overall Survival in Patients with Metastatic Oesophageal or Gastric Cancer. Cancers 11 (2019).
47. Sivendran, S., Latif, A., McBride, R. B., *et al.* Adverse event reporting in cancer clinical trial publications. Journal of Clinical Oncology 32, 83–89 (2014).
48. Van de Water, L. F., van Kleef, J. J., Dijksterhuis, W. P., *et al.* Communicating treatment risks and benefits to cancer patients: a systematic review of communication methods. Quality of Life Research 29, 1747–1766 (2020).
49. Nielsen, J. Enhancing the explanatory power of usability heuristics. Proceedings of the SIGCHI conference on Human Factors in Computing Systems, 152–158 (1994).
50. Sommerville, I. Software engineering (10th edition), 811 (Pearson Education Limited. Essex, England, 2016).
51. Hakone, A., Harrison, L., Ottley, A., *et al.* PROACT: iterative design of a patient-centered visualization for effective prostate cancer health risk communication. IEEE transactions on visualization and computer graphics 23, 601–610 (2016).
52. Brooke, J. SUS: a "quick and dirty'usability. Usability evaluation in industry 189 (1996).
53. Clegg, D. & Barker, R. Case Method Fast-Track: A Rad Approach (Computer Aided System Engineering). Project Smart, 207 (1994).
54. Villanueva, R. A. M., Chen, Z. J. & Wickham, H. ggplot2: Elegant graphics for data analysis using the grammar of graphics, 160–167 (Springer-Verlag, New York, NY, 2016).

55. RStudio. shiny: web application framework for R. 2013.
56. Onwuegbuzie, A. J., Dickinson, W. B., Leech, N. L. & Zoran, A. G. A qualitative framework for collecting and analyzing data in focus group research. International journal of qualitative methods 8, 1–21 (2009).
57. Aranyi, G., Van Schaik, P. & Barker, P. Think-aloud Demonstration Video. English. 2012. `https://www.youtube.com/watch?v=BwpPIiBK0cA`.
58. Van Merriënboer, J. J. G. Ten Steps to Complex Learning (Citeseer, 2007).
59. Van Merriënboer, J. J. G. Alternate models of instructional design: Holistic design approaches and complex learning. Trends and issues in instructional design and technology, 72–81 (2007).
60. Van Merriënboer, J. J. & Kirschner, P. A. Three worlds of instructional design: State of the art and future directions. Instructional Science 29, 429–441 (2001).
61. Henselmans, I., Smets, E. M. A., de Haes, J., *et al.* A randomized controlled trial of a skills training for oncologists and a communication aid for patients to stimulate shared decision making about palliative systemic treatment (CHOICE): study protocol. BMC cancer 18, 1–13 (2018).
62. Gilligan, T., Coyle, N., Frankel, R. M., *et al.* Patient-clinician communication: American Society of Clinical Oncology consensus guideline. Obstetrical & Gynecological Survey 73, 96–97 (2018).
63. Kiely, B. E., Stockler, M. R. & Tattersall, M. H. N. Thinking and talking about life expectancy in incurable cancer. Seminars in oncology 38, 380–385 (2011).
64. Kiely, B. E., Soon, Y. Y., Tattersall, M. H. N. & Stockler, M. R. How long have I got? Estimating typical, best-case, and worst-case scenarios for patients starting first-line chemotherapy for metastatic breast cancer: a systematic review of recent randomized trials. Journal of Clinical Oncology 29, 456–463 (2011).
65. Zipkin, D. A., Umscheid, C. A. & Keating, N. L. Evidence-based risk communication: a systematic review. Ann Intern Med 161, 270–280 (2014).
66. Sella, T., Botser, D., Navon, R., *et al.* Preferences for disclosure of disease related information among thoracic cancer patients. Lung Cancer 88, 100–103 (2015).
67. Hoffmann, T. C. & Del Mar, C. Patients' expectations of the benefits and harms of treatments, screening, and tests: a systematic review. JAMA internal medicine 175, 274–286 (2015).
68. Henselmans, I., van Laarhoven, H. W. M., de Haes, H. C. J. M., *et al.* Training for medical oncologists on shared decision-making about palliative chemotherapy: A randomized controlled trial. The oncologist 24, 259 (2019).
69. Colevas, A. D. & Setser, A. The NCI Common Terminology Criteria for Adverse Events (CTCAE) v 3.0 is the new standard for oncology clinical trials. Journal of Clinical Oncology 22, 6098–6098 (2004).
70. Scott, N., Fayers, P., Aaronson, N., *et al.* Reference Values Manual (2008).
71. Stiefel, F., Barth, J., Bensing, J., *et al.* Communication skills training in oncology: a position paper based on a consensus meeting among European experts in 2009. Annals of oncology 21, 204–207 (2010).
72. De Vries, A. M. M., De Roten, Y., Meystre, C., *et al.* Clinician characteristics, communication, and patient outcome in oncology: a systematic review. Psycho-Oncology 23, 375–381 (2014).
73. Grossman, R. & Salas, E. The transfer of training: what really matters. International Journal of Training and Development 15, 103–120 (2011).
74. Cepeda, N. J., Coburn, N., Rohrer, D., *et al.* Optimizing distributed practice: theoretical analysis and practical implications. Experimental psychology 56, 236 (2009).

75. Dempster, F. N. Spacing effects and their implications for theory and practice. Educational Psychology Review 1, 309–330 (1989).
76. Dempster, F. N. & Perkins, P. G. Revitalizing classroom assessment: Using tests to promote learning. Journal of Instructional Psychology 20, 197 (1993).
77. Donovan, J. J. & Radosevich, D. J. A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. Journal of Applied Psychology 84, 795 (1999).
78. Moss, V. D. The efficacy of massed versus distributed practice as a function of desired learning outcomes and grade level of the student. PhD thesis (1995).
79. Janiszewski, C., Noel, H. & Sawyer, A. G. A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. Journal of consumer research 30, 138–149 (2003).
80. Van Kleef, J. J., van den Boorn, H. G., Verhoeven, R. H., *et al.* External validation of the dutch SOURCE survival prediction model in belgian metastatic oesophageal and gastric cancer patients. Cancers 12 (2020).
81. Hillen, M. A., Gutheil, C. M., Smets, E. M. A., *et al.* The evolution of uncertainty in second opinions about prostate cancer treatment. Health Expectations 20, 1264–1274 (2017).
82. Van den Boorn, H. G., Stroes, C. I., Zwinderman, A. H., *et al.* Health-related quality of life in curatively-treated patients with esophageal or gastric cancer: A systematic review and meta-analysis. Critical Reviews in Oncology/Hematology 154, 103069 (2020).
83. Van den Ende, T., Abe Nijenhuis, F. A., van den Boorn, H. G., *et al.* COMplot, A Graphical Presentation of Complication Profiles and Adverse Effects for the Curative Treatment of Gastric Cancer: A Systematic Review and Meta-Analysis. Frontiers in Oncology 9, 1–12 (2019).
84. Source web-interface. `https://source.amsterdamumc.org` (2021).

Supplement

Comment	Support	Result
Patients		
Pictograph legend should be on top.	3 out of 6	Pictograph legend is moved to top of graph.
Title is unclear when showing prediction for a certain month/year.	3 out of 6	The title was changed to display e.g. 6 months instead of specific date.
Pictograph is clearer than line graph.	6 out of 6	The pictograph is shown by default, the line graph is optional.
The distance between bars in barchart is too large.	4 out of 6	The distance between bars is made smaller.
Linegraph is clearer than barchart to display HRQoL.	4 out of 6	HRQoL predictions are now only shown in a line graph.
Unclear what mild/severe toxicity means.	4 out of 6	In the training focus is given to explain this clearly to patients.
Don't use abbreviations of chemotherapy names.	1 out of 6	Due to low support among patients, complex chemotherapy names and use of abbreviations in clinical practice, the abbreviations remain in use.
The text must be larger and clearer.	3 out of 6	The font and font size was changed.
The meaning of the pictograph is unclear.	2 out of 6	In the training focus is given to explain the pictographs clearly to patients.
Scenario function (best-case, typical outcome and worst-case) is difficult to understand.	3 out of 6	In the training focus is given to explain this clearly to patients.
Scenario function is helpful/important and relevant.	4 out of 6	The scenario function remains in the tool and the training for physicians focuses on explaining the functionality.
Bars in the toxicity graph displaying mild and severe toxicity are interpreted as both stacked and non-stacked.	5 out of 6	The bars displaying mild and severe toxicity are placed next to each other with 90% overlap to clarify that the bars are not stacked.
The difference between mild and severe toxicity is too small.	3 out of 6	The severe toxicity bars are made darker to increase contrast with mild toxicity.
The main application colour is green which is not preferred.	3 out of 6	The main application colour is changed to a more neutral blue colour.
Physicians		
The input of the models requires a specific diagnosis date which is impractical.	3 out of 3	The date input is removed from the tool and predictions are not shown for a specific date but rather in the form '6 months'.
The label 'Oesophagus NOS' is unclear.	1 out of 3	The label is changed to 'Other'.
The label 'lymph node metastasis' is unclear.	1 out of 3	The label is changed to 'local lymph node metastasis'.
Choosing the exact tumour location can be difficult in practice.	1 out of 3	The descriptions of the tumour location have been changed and a visualisation of the stomach and oesophagus have been added for extra clarity.
There should be more input values on a single page.	1 out of 3	The number of inputs on a page has increased to fill an entire screen.
Show the survival in text next to the pictograph.	1 out of 3	The survival has been added to the legend.
Cannot find the function for most probable survival.	3 out of 3	The function is now displayed in text instead of an icon and is clearly stated under diagram options.

Supplementary Table 1 – *continued from previous page*

Comment	Support	Result
Missing median time when hovering over graph.	2 out of 3	When hovering over the graph a dotted line is shown to indicate the time on the x-axis.
Bars in the side effects graph displaying mild and severe side effects are interpreted as both stacked and non-stacked.	1 out of 3	The bars displaying mild and severe side effects are placed next to each other with over 90% overlap to clarify that the bars are not stacked.
Set side effects in order of decreasing occurrence.	2 out of 3	This suggestion was applied to the tool. However, it was found that it was confusing that a side effects would shift in the bar-graph with the addition of another treatment. All side effects now have a set order.
Pop-up when no chemo is selected can be annoying.	2 out of 3	The pop-up is removed.
The absence of a save button after filling in the models input values is confusing.	2 out of 4	A save button is added.
The side effects graph is difficult to understand without extra explanation.	1 out of 4	A clarifying picture is added as a legend to indicate which bars describe mild side effects and which are severe side effects.
The x-axis of the scenario graph is hard to read.	1 out of 4	A hover function is added with a vertical line to the x-axis, in the future x-axis values are also displayed in the hover window.
Display the x-axis of the survival line graph in half year increments instead of two months increments.	1 out of 4	When the x-axis range is set to 1 year, the x-axis is displayed in increments of one month. When the range extends 1 year, increments of half a year are displayed.
Experts		
The direct appearance of the survival graph after filling in the input values can be confronting for patients.		After inputting the patient data, an empty screen is shown after which the physician can select an outcome to display.
Display the number of coloured icons in the pictograph in numbers too, above/on the side of the graph.		The number of coloured icons is displayed in the legend above the graph.
Display the icons in the survival icon array as overlapping when this is meant.		When two treatments are compared, colouring an icon in two colours was found to be confusing, so the two treatments remain in a single pictograph. However, for three or more treatments, each treatment is displayed in a single pictograph.
Display the scale on the y-axis of the survival line graph as frequencies (100/100) instead of percentages (100%).		The suggested change was judged to make reading the graph confusing for persons with high graph literacy (e.g. physicians) and therefore not applied.
Accompanying verbal explanation of the scenario's can be difficult for the physician to come up with him-/herself.		An information button will be added to the scenario's graph.
The reference score in the health-related quality of life graph is not displayed in the legend.		A label is added to the side of the graph.
The side effects graph is hard to read: it is unclear whether the bars are stacked or not.		The display is changed from only side-to-side bars to overlapping bars, for the severe and mild side effects, to clarify that the bars are not stacked.

Supplementary Table 1 – *continued from previous page*

Comment	Support	Result
The automatically generated texts for the summary function are difficult to understand for low literacy patients due to terminology.		Difficult terminology is replaced by easy to understand alternatives where possible. Abstract concepts (i.e. health-related quality of life) are explained in the summary text.
The summary function should have display options of its own.		Summary display options are added, but turn out to make the summary page more unclear by an overload of options. Also, the website speed is decreased a lot by this added functionality. The options are therefore removed and the preferences as set in the main graph on the outcome tab are also displayed in the summary.

Supplementary Table 1: Detailed adaptations to the SOURCE tool during the modelling phase.

	Recommendation
1.	Do not use only general verbal descriptions (e.g. 'often', 'some people'), but add numbers (e.g. 50% or 50/100) to your description of the risk or benefit. Solely verbal descriptions might easily be misunderstood or misjudged by patients.
2.	Accompany your information with visual representations of the information to facilitate the understanding, for instance by using the SOURCE tool.
3.	Do not use relative risk reductions. Research shows that these are often more often misunderstood than absolute risk reductions and that relative risk reductions can influence treatment choice.
4.	Preferably use frequencies and/or percentages to inform patients on risks and benefits of treatment. Be sure to use clear descriptions that cannot be easily misunderstood.
5.	Add a reference class to your frequency/percentage (time frame, place, total number of people, etc.). To achieve this ask yourself the following questions: a. Whom does this number apply to? b. Which period of time or moment in time does this number apply to?
6.	Use the three scenarios as introduced by Kiely et al.[63,64] to inform about survival. These scenario's stress the range of outcomes allowing patients to prepare for the worst and hope for the best.

Supplementary Table 2: Recommendations on risk communication displayed in the e-learning training.

Chapter 11

General discussion

General discussion

Patients with oesophageal and gastric cancer and their physicians often face the difficult challenge to determine which course of treatment is best for them. This decision is hampered by the uncertainty of the treatment outcomes; i.e. how life will look like after completion of treatment, how long the patient may live and how the treatment will influence the patient's health-related quality of life (HRQoL). The main aim of this thesis is to develop prediction models and perform statistical analyses to provide evidence-based information, that is as accurate as possible, to reduce this uncertainty of treatment outcomes. With more accurate information it is possible for patients and their physicians to determine which course of treatment coincides best with the patient's treatment goals.
In this chapter, the results of the previous chapters are discussed, in the context of four interrelated topics.

The use of meta-analyses versus individual patient data for the prediction of treatment outcomes

The analyses in this thesis are based on two main sources of data: meta-analyses where data from individual studies are combined, and population-based registries containing individual patient data. While both sources contain important and relevant information regarding treatment outcomes, the sources differ in the information they provide, the generalisability of the results, and their advantages and disadvantages.

The meta-analyses reported in this thesis investigate the benefits and harms of specific treatment interventions. While randomised controlled trials (RCTs) serve as the 'gold standard' in medical science[1], they often have narrowly-defined inclusion criteria, which hamper generalisation of the results to the entire population of patients. This form of selection bias is illustrated in RCTs where often only patients who are in relatively good health are included.[2] In these cases it is unknown whether the found treatment outcomes will also be observed in patient groups with poorer health. Another drawback in the use of meta-analyses is the lack of individual patient data (IPD).[3] Individual studies provide summary information gathered on the entire sample and certain subgroups, but information on an individual level is generally not provided.[4] In these cases, authors of the studies have to be contacted to share this information.[3] As was demonstrated for the prediction models in Chapters 3–6, IPD enables the prediction of treatment outcomes in individuals and may lead to more targeted predictions as multiple variables influence the outcome and are

corrected for one another in the models. However, sharing of data may face extra barriers. While steps can be undertaken to ensure privacy regulations, authors are often reluctant to share the IPD, and data sharing may be burdensome and/or costly.[5] For example, authors may plan to use the data for future publications and therefore may not wish to share the data prior to that date. Data sharing requirements from journals[6] may aid in this process. Moreover, methods such as secure multi-party computation using encryption, can be used to ensure that others cannot access IPD directly, while at the same time it is possible to run analyses on IPD from these sources.[7] Although promising, this technique is still in its infancy and currently is not widely used. Another issue to consider is the quality of individual studies used in meta-analyses. Although standards exist to decrease bias in reporting[8], the results in Chapters 2, 7 and 8 indicate that in general, the quality of included studies is poor. It is important and often difficult to ensure reporting consistency among studies when performing a meta-analysis.
While meta-analyses can pose obstacles to providing predictions of treatment outcomes, there are also advantages. For example, included studies are available from a wide variety of sources such as RCTs, cohort studies or case-series and are less sensitive to outliers in single studies.[9] Another important advantage of meta-analyses is the overall number of patients that are included. This plays an important role, especially for secondary outcomes such as HRQoL which are often not routinely gathered in all patients.

Population-based registries, on the other hand, reflect information obtained from daily clinical practice and often include IPD. The main advantage is that they minimise selection bias by including all patients in a region during a certain time frame. In this way it is possible to collect a large amount of IPD from a heterogenous group of patients and develop a prediction model for that population. With the use of data of patient, tumour and treatment characteristics it is also possible to create individualised predictions, which may agree more with the observed outcome of a particular patient or groups of patients.

In general, IPD based on clinical practice is therefore the preferable source of data to develop clinical prediction models, as the models may provide more accurate and personalised predictions on treatment outcomes. In cases where the registries have not included a large number of patients, meta-analyses can provide valuable, additional information and predictions can become more precise due to the large numbers of included patients. This is exemplified by the prediction model for HRQoL presented in Chapter 9, for which insuffi-

cient data for patients with metastatic disease was available. As this model had a relatively poor predictive performance, the meta-analyses described in Chapters 7 and 8 can provide valuable information by establishing and providing insight into global trends in HRQoL in oesophagogastric cancer patients.

Better models through more and better data

One of the most important aspects of creating clinical prediction models is the availability of large datasets and informative variables that relate to the outcome.[10] A major problem in the development of prediction models is model overfitting, where the prediction model mainly learns patterns in the training dataset that do not extrapolate to future observations.[10] The primary causes for overfitting include the use of small data sets for model development[11] and, relatedly, the use of too many predictors in the model.[12] With larger datasets, patterns in data, such as relationships between variables and outcome, can be determined more reliably.[13] Furthermore, the patterns generalise better to external data sets, increasing the model's performance. Likewise, the models' performance can also be determined more reliably as the variability of the performance measures decreases in larger data sets.[10] The reduction of model overfitting in larger dataset allows for models to increase the number of parameters. This can lead in turn to better performance and the use of 'big data' algorithms which rely on large numbers of parameters, such as artificial neural networks.[14] These algorithms are able to establish more complex relations between predictors and outcome and, given a sufficient amount of data, are able to achieve a high performance. These algorithms also make fewer assumptions (such as linearity or additivity) on the underlying data. Overall, it can thus be stated that larger datasets can lead to prediction models with higher performance, which are less susceptible to overfitting.

Whereas expansion of the datasets is therefore recommended, it is not always achievable in practice. In Chapters 3, 5 and 6, for example, the Netherlands Cancer Registry was used to create prediction models for overall survival. This population-based registry already covers the entire Dutch population, so extension of this dataset will involve for instance the inclusion of data from earlier diagnosis years or the use of external cancer registries. Using earlier diagnosis years could lead to suboptimal models, as the influence of variables such as treatment on overall survival will change over the years.[15] Usage of foreign registry data may be difficult to achieve as variables in other registries may be defined differently. Furthermore, model parameters are often different between countries as, for example, treatment protocols are different.

Both obstacles have been addressed in Chapter 4. In the prediction model for HRQoL (Chapter 9), expansion of the dataset is necessary as a relatively small number of patients were included in this voluntary study. This could be improved by promoting the use of HRQoL questionnaires at diagnosis and after treatment completion.

A further improvement to the prediction models is the expansion of the variables that are collected for each patient. This is exemplified in the prediction model for HRQoL (Chapter 9), where the model performance was low due to the lack of predictive variables. Further research into predictive variables and inclusion of such variables, for instance illness perception[16], could lead to more accurate predictions. With more relevant variables that relate to the outcome variable, models can be based on a richer dataset that allows for better performant models. In the future, additional sources of data can be investigated and incorporated in these datasets. For instance, WHO performance status[17] (Chapters 5 and 6) is an important predictor of overall survival. This nominal variable consists of five levels of physical activity, from unrestricted normal activity to confinement to bed or chair. Establishing which level is most relevant for a patient may be arbitrary and imprecise. A more reliable measure of physical performance may be provided by wearables such as smartwatches, which can also monitor physical performance over time.[18,19] Usage of such (external) data sources may become more relevant given the increasing use of such wearables. Also the use of natural language processing could play an important role in the extraction of additional valuable information from patient records.[20] This currently is a labour-intensive task, and the automatic extraction of information can aid in creating larger datasets, with more information available per patient.

Through the inclusion of more data per patient, variables can be measured more accurately and more relevant predictors are included in the model. And with the inclusion of data of more patients, the influence of predictors on outcomes can be established more accurately. Altogether, dataset extension can therefore lead to more accurate prediction models in the future.

The use of prediction models and meta-analyses in clinical practice

The models and meta-analyses presented in this thesis are devised to aid in the complex task of accurately informing patients on treatment outcomes, which is especially important when it influences treatment decisions.[21] Evidence-based medicine requires the use of the latest research findings in clinical practice.[22] In the context of this thesis, the goal of evidence-based medicine is to come to a more objective and more accurate description of treatment outcomes.[23] Without the use of prediction models, clinicians rely on their education, expert knowledge and clinical experience. A systematic review analysing the survival predictions made by physicians in terminally-ill cancer patients demonstrated that the physicians' predictions often were inaccurate.[24] The physicians' predictions were in general overoptimistic, although highly correlated with observed survival.[24] This study suggests the need for clinical prediction models to aid in more accurately predicting outcomes such as survival. Indeed, the SOURCE prediction models (Chapters 3, 5 and 6) show a high discriminative ability and close calibration of predicted and observed survival during validation. Prediction models taking multiple variables into account and being less subjective may improve the predictive accuracy of physicians[25] and therefore may be of added benefit in clinical practice.

A web-interface, as presented in Chapter 10, can be used to present prediction models and results of meta-analyses in a user-friendly manner.[26] In developing the SOURCE web-interface, we aimed to implement the models and findings of the meta-analyses in clinical practice and engage patients and physicians in shared decision making.[27,28] The purpose of this web-interface is to support the clinical consultation between a physician and their patient, and therefore it is advised that patients do not use the prediction models without their physician. The physician can choose and interpret the relevant predictors, specify which treatments are possible for the patients, and interpret the outcomes.

Although prediction models and meta-analyses are intended to provide more accurate and targeted information, they may give incorrect information when not considering the dataset's case mix.[29,30] As discussed in the context of the first topic, meta-analyses may suffer from selection bias and outcomes reported by these analyses may differ from those in clinical practice. While prediction models based on population-based registries suffer less from selection bias, one should always consider the relevance and the case mix of models for individual patients.[23] For instance, if a patient suffers from severe

obesity and differs in that sense from the patients used to develop the prediction model, the model may provide too optimistic predictions for that patient. This lack of generalisability has also been shown in Chapter 4, where a prediction model for overall survival in oesophageal cancer patients did not transfer to an external cohort. It is therefore important to consider whether the patient for whom a prediction is made, is similar to the patients on whom the analyses are based, and that the predictors (such as proposed treatment) are relevant for that patient.[29] The relatedness between the training and validation datasets also needs be taken into account to interpret model performance.[31] As discussed in Chapter 10, the source and limitations of the prediction models and meta-analyses are vital to address when training physicians to use prediction models and interpret their results.

In the end, clinical prediction models and meta-analyses can be complementary to the knowledge and clinical experience of physicians. By displaying the models and their outcomes in a user-friendly manner, while also training physicians to use these models by instructing them how to interpret the predictions and communicate these effectively to patients, a synergy can arise between the evidence-based predictions provided by the models and the physicians' expertise and experience.

Requirements for a successful implementation of prediction models in clinical practice

The successful implementation of prediction models in clinical practice can be hard to achieve and depends on a number of aspects.

First of all, the model's performance has to be well-established and verified. As discussed in Chapter 2, the performance of prediction models in subsequent validation studies often is significantly lower than in the development studies.[12] While techniques such as predictor selection[32] and ensemble learning[33] can be employed to increase the model's performance, measuring whether the model performs worse on unseen data (thus manifesting overfitting on the developmental dataset) is essential. Cross-validation was employed to avoid overfitting by obtaining unbiased performance estimates in the models discussed in this thesis. Moreover, the models' performance was measured similarly to how the models are used in practice.[10] External validation, whether it is employed in an external population as in Chapter 4, or in later years in the same population (internal-external temporal cross-validation[30]), is key to establishing the models' performance. Without such validation, the model's performance should be viewed with scepticism.

Model updating is needed to maintain the models' performance.[10,34] Populations on which the models are based change over time, as discussed earlier.[15] Also the effects of treatment on outcomes, such as survival, change over time, as treatment effectiveness may increase over time.[35] Model updating with more recent data capturing these changes should maintain the accuracy of the predictions and possibly even improve it.[36] The SOURCE prediction models for survival particularly need updating, since we used internal-external temporal cross-validation to measure model performance. In these studies, multiple prediction models were made and evaluated on data of the following year. It therefore remains uncertain what the performance of the models is when used in subsequent years. Model updating therefore is an important aspect in determining and maintaining the performance of the models in practice.

As indicated in Chapter 10, the limitations and methods used to create prediction models are critical to assess and to interpret their predictions. This can be a challenging task as prediction models do not offer a causal explanation.[37] For physicians with no prior experience with machine learning, prediction models are therefore sometimes seen as 'black boxes', where data is inputted and, after a 'hidden' process, a prediction is made.[38] Objections to using these models are understandable, especially as they may influence treatment choices.[39] The use of relatively 'simple' algorithms that have interpretable model parameters, such as linear models or Cox regression, can aid in keeping the models understandable. These algorithms, however, cannot be used in all models due to the complex nature of the prediction problem, e.g. image recognition to detect tumours in MRI scans.[40] Explainable artificial intelligence (XAI) focuses on making such complex models more transparent, interpretable and explainable.[41] This is achieved by explaining the technique used to create the model, clarifying the validation process and, more importantly, indicate how models arrive at their specific predictions.[41] This can be achieved, for example, by determining which patient, tumour and treatment variables were pivotal in predicting the outcome and by visualising how the variables influenced the predictions.[42] These techniques can be used to explain, for example, how the prediction model for HRQoL (Chapter 9) arrives at its predictions more intuitively. Also, involving end-users of the models (patients and physicians) in early stages may point out issues in the prediction model, which may lead to improvements and increased trust. The evaluation of the prediction models in practice may also increase the trust in the models. The web-interface and accompanying training for physicians, reported in Chapter 10, are currently evaluated in a stepped-wedge trial (registered under NCT04232735, the SOURCE trial). Through this trial, the effect on infor-

mation provision can be measured and with the use of the models in clinical practice, we also aim to increase the trust in the models. Together, these techniques can be used effectively to create better explainable and more trusted prediction models, and may aid implementation and use of the models in practice.

Concluding remarks

To conclude, this thesis describes analyses aimed at predicting treatment outcomes in patients with oesophagogastric cancer. Prediction models based on individual patient data are preferred to meta-analyses as they allow for more personalised predictions, yet meta-analyses can provide useful overviews of patterns in large populations. In the future, prediction models can be improved by expanding data sets by including more patients and/or more relevant variables. Establishing the models' quality and limitations is essential. The resulting prediction models and analyses can provide complementary information and visualisations to facilitate shared decision making in clinical practice.

References

1. Stewart, L. A. & Michael, J. C. Practical methodology of meta-analyses (overviews) using updated individual patient data. Statistics in Medicine 14, 2057–2079 (1995).
2. Peppercorn, J. M., Weeks, J. C., Cook, E. F. & Joffe, S. Comparison of outcomes in cancer patients treated within and outside clinical trials: Conceptual framework and structured review. Lancet 363, 263–270 (2004).
3. Veroniki, A. A., Straus, S. E., Ashoor, H., *et al.* Contacting authors to retrieve individual patient data: Study protocol for a randomized controlled trial. Trials 17, 1–8 (2016).
4. Lundh, A., Krogsbøll, L. T. & Gøtzsche, P. C. Access to data in industry-sponsored trials. The Lancet 378, 1995–1996 (2011).
5. Wicherts, J. M., Borsboom, D., Kats, J. & Molenaar, D. The poor availability of psychological research data for reanalysis. American Psychologist 61, 726–728 (2006).
6. Taichman, D. B., Sahni, P., Pinborg, A., *et al.* Data Sharing Statements for Clinical Trials A Requirement of the International Committee of Medical Journal Editors. JAMA - Journal of the American Medical Association 317, 2491–2492 (2017).
7. Goldreich, O. Secure Multi-Party Computation (1998).
8. Schulz, K. F., Altman, D. G. & Moher, D. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials. Annals of Internal Medicine 152, 726–732 (2010).
9. Rothman, K., Greenland, S. & Lash, T. Modern epidemiology (2008).
10. Steyerberg, E. Clinical Prediction Models (Springer, New York, NY, 2009).
11. Ying, X. An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series 1168, 022022 (2019).
12. Ruppert, D. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Journal of the American Statistical Association 99, 567–567 (2004).
13. Steyerberg, E. W., Eijkemans, M. J., Harrell, F. E. & Habbema, J. D. F. Prognostic modelling with logistic regression analysis: A comparison of selection and estimation methods in small data sets. Statistics in Medicine 19, 1059–1079 (2000).
14. Sagiroglu, S. & Sinanc, D. Big data: A review. Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013, 42–47 (2013).
15. Bosetti, C., Levi, F., Ferlay, J., *et al.* Trends in oesophageal cancer incidence and mortality in Europe. International Journal of Cancer 122, 1118–1129 (2008).
16. Ashley, L., Marti, J., Jones, H., *et al.* Illness perceptions within 6 months of cancer diagnosis are an independent prospective predictor of health-related quality of life 15 months post-diagnosis. Psycho-Oncology 24, 1463–1470 (2015).
17. Yates, J. W., Chalmer, B. & McKegney, F. P. Evaluation of patients with advanced cancer using the karnofsky performance status. Cancer 45, 2220–2224 (1980).
18. Nyrop, K. A., Deal, A. M., Choi, S. K., *et al.* Measuring and understanding adherence in a home-based exercise intervention during chemotherapy for early breast cancer. Breast Cancer Research and Treatment 168, 43–55 (2018).
19. Low, C. A., Dey, A. K., Ferreira, D., *et al.* Estimation of symptom severity during chemotherapy from passively sensed data: Exploratory study. Journal of Medical Internet Research 19, e9046 (2017).
20. Liu, F., Weng, C. & Yu, H. Natural Language Processing, Electronic Health Records, and Clinical Research. Clinical Research Informatics, 293–310 (2012).
21. Stiggelbout, A. M., Van der Weijden, T., De Wit, M. P., *et al.* Shared decision making: Really putting patients at the centre of healthcare. The BMJ (Online) 344 (2012).

22. Sur, R. L. & Dahm, P. History of evidence-based medicine. Indian Journal of Urology 27, 487–489 (2011).
23. Steyerberg, E. W. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2009.
24. Glare, P., Virik, K., Jones, M., *et al.* A systematic review of physicians' survival predictions in terminally ill cancer patients. The BMJ 327, 195–198 (2003).
25. Vickers, A. J. Prediction models in cancer care. CA: A Cancer Journal for Clinicians 61 (2011).
26. Bonnett, L. J., Snell, K. I., Collins, G. S. & Riley, R. D. Guide to presenting clinical prediction models for use in clinical settings. The BMJ (Online) 365 (2019).
27. Politi, M. C., Han, P. K. & Col, N. F. Communicating the uncertainty of harms and benefits of medical interventions. Medical Decision Making 27, 681–695 (2007).
28. Hodgkinson, K., Butow, P., Hunt, G. E., *et al.* The development and evaluation of a measure to assess cancer survivors' unmet supportive care needs: TheCaSUN (Cancer Survivors' Unmet Needs measure). Psycho-Oncology 16, 796–804 (2007).
29. Kappen, T. H., Vergouwe, Y., van Klei, W. A., *et al.* Adaptation of Clinical Prediction Models for Application in Local Settings. Medical Decision Making 32 (2012).
30. Steyerberg, E. W. & Harrell, F. E. Prediction models need appropriate internal, internal-external, and external validation. Journal of Clinical Epidemiology 69, 245–247 (2016).
31. Debray, T. P., 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).
32. Hawkins, D. M. The Problem of Overfitting. Journal of Chemical Information and Computer Sciences 44, 1–12 (2004).
33. Sollich, P. & Krogh, A. Learning with ensembles: how over-fitting can be useful (1995).
34. Davis, S. E., Lasko, T. A., Chen, G. & Matheny, M. E. Calibration Drift Among Regression and Machine Learning Models for Hospital Mortality. AMIA Annual Symposium proceedings. AMIA Symposium 2017, 625–634 (2017).
35. De Vos-Geelen, J., Geurts, S. M., Valkenburg-Van Iersel, L. B., *et al.* Trends in treatment and overall survival among patients with proximal esophageal cancer. World Journal of Gastroenterology 25, 6835–6846 (2019).
36. Minne, L., Eslami, S., de Keizer, N., *et al.* Statistical process control for monitoring standardized mortality ratios of a classification tree model. Methods of Information in Medicine 51, 353–358 (2012).
37. Shmueli, G. To explain or to predict? Statistical Science 25, 289–310 (2010).
38. Nicholson Price, W. Big data and black-box medical algorithms. Science Translational Medicine 10 (2018).
39. Gianfrancesco, M. A., Tamang, S., Yazdany, J. & Schmajuk, G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Internal Medicine 178, 1544–1547 (2018).
40. Kavitha, A. R. & Chellamuthu, C. Detection of brain tumour from MRI image using modified region growing and neural network. Imaging Science Journal 61, 556–567 (2013).
41. Roscher, R., Bohn, B., Duarte, M. F. & Garcke, J. Explainable Machine Learning for Scientific Insights and Discoveries. IEEE Access 8, 42200–42216 (2020).
42. Samek, W., Wiegand, T. & Müller, K.-R. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models (2017).

Chapter 12

Summary
Samenvatting

Summary

Various treatments exist to remove tumours, alleviate symptoms and combat recurrence in oesophagogastric cancer patients. Survival often remains poor and treatments can come at the cost of health-related quality of life (HRQoL). Providing accurate information to patients and physicians about the outcomes of relevant treatments, such as survival and HRQoL, is crucial to determine which course of treatment is best and coincides with the patients' preferences. However, these outcomes depend on many factors and accurately determining these outcomes beforehand is complex. The aim of this thesis is therefore to provide evidence-based information on treatment outcomes that is as accurate as possible, by performing meta-analyses and developing prediction models for patients with oesophagogastric cancer.
Chapters 1 and 11 of this thesis provide a general introduction and general discussion, respectively. Below is a summary of the other chapters in this thesis.

Chapter 2 provides a systematic review and meta-analysis of prediction models that have been developed to predict treatment outcomes in patients with oesophagogastric cancer. Prediction models often are evaluated with the c-index and model calibration. The c-index states the model's discriminative ability, and typically ranges from 0.5 (no discrimination at all) to 1 (perfect discrimination). The model calibration describes the agreement between the predicted and observed outcomes, and has an intercept of 0 and slope of 1 when the predictions are perfect. A total of 47 prediction models were found, described in 45 development and 16 validation studies. A meta-analysis of the models' c-indices indicate that discriminatory ability of these models is fair with a value of 0.75 (range: 0.65–0.85), and that the c-index in validation studies is lower than during model development (0.73 vs. 0.76, $p = 0.01$). Most models predict survival after a curative resection and there are no models that predict HRQoL. There is a need for externally-validated models that provide information on both the benefits and harms for a variety of treatments.

In Chapter 3 we report on the development and validation of prediction models for overall survival in metastatic oesophagogastric cancer patients. Two multivariate Cox regression models were developed with data from respectively 8,010 oesophageal and 4,763 gastric cancer patients with metastases, diagnosed between 2005–2015 in The Netherlands. Predictor selection was performed via a Delphi consensus and the Akaike Information Criterion. The models were evaluated through internal-external temporal cross-validation

and demonstrated a good calibration (intercept: 0.00, slope: 1.00). The c-index was 0.71 (95% CI: 0.70 – 0.71) in the oesophageal model and in the gastric model the c-index was 0.68 (95% CI: 0.67 – 0.69). The models have fair c-indices and can support shared decision making by providing personalised survival estimates.

Chapter 4 reports on the external validation of the two prediction models for overall survival developed in Chapter 3. In this study, the models were validated with a Belgian cohort of 2,514 metastatic oesophageal cancer patients, and 1,583 metastatic gastric cancer patients. The Belgian cohort differed in multiple patient, tumour and treatment characteristics from the Dutch cohort. The gastric model showed in this validation study a similar calibration to the development study (intercept: 0.02, slope: 0.91), as well as a similar c-index of 0.66. The oesophageal model, however, had a poor calibration (intercept: 0.30, slope: 0.42) and a lower c-index of 0.64. These findings demonstrate that the oesophageal model did not transport well to the Belgian population, and that future models should reduce the number of parameters to arm against model overfitting.

Chapter 5 describes the development and validation of a set of four prediction models, aimed at predicting overall survival in both metastatic and potentially curable oesophageal and gastric cancer patients. The models provide an update to the models reported in Chapter 3, and an extension of the models to potentially curable patients. These Cox regression models are based on a Dutch cohort (N = 13,365) diagnosed between 2015–2018, and were evaluated through internal-external temporal cross-validation. The c-indices for metastatic disease are 0.72 (95% CI: 0.71–0.74) and 0.73 (95% CI: 0.69–0.75), respectively, for oesophageal and gastric cancer. The models for potentially curable oesophageal and gastric cancer patients have c-indices of 0.80 (95% CI: 0.75–0.84) and 0.78 (95% CI: 0.74–0.82), respectively. All models demonstrated an overall good calibration. The models can be used in both metastatic and potentially curable disease and are the first in oesophagogastric cancer to predict overall survival for a variety of treatments. Further research is needed to demonstrate the added value of the use of these models in clinical practice.

In Chapter 6 we report on the development and validation of a prediction model for overall survival in patients with metastatic pancreatic ductal adenocarcinoma. A Cox regression model was developed using methodology that resembles the one described in Chapter 5. The model was based on a Dutch cohort of 4,739 patients diagnosed between 2015–2018. An internal-external temporal cross-validation scheme was used to determine the model's performance. The model demonstrates an overall good calibration, and a c-index of 0.72 (95% CI: 0.71 – 0.73). This study validates the methodology used in previous chapters, and shows that a similar model performance is obtained when the same steps are taken to develop and evaluate the model in a different setting.

In Chapter 7 we report on a systematic review and meta-analysis of HRQoL in oesophageal and gastric cancer patients undergoing palliative systemic therapy. A total of 43 randomised controlled trials, with a total of 13,727 patients, were included. HRQoL scores were meta-analysed using linear mixed-effects models with cubic splines to capture non-linear effects. The analysis demonstrates that patients have an impaired HRQoL before the start of treatment, with a EORTC global health status score of 54.6 (95% CI: 51.9–57.3) in first-line treatments and 57.9 (95% CI: 55.7–60.1) in beyond first-line treatments. An analysis of the change in HRQoL over time demonstrates that HRQoL remained stable for most treatments. In HER2-negative cancers, anthracycline-based triplets and fluoropyrimidine-based doublets without cisplatin show better HRQoL in first-line treatments compared to other first-line treatments. In beyond first-line treatments, taxanes and targeted agents were shown to provide increased HRQoL compared to best supportive care.

A systematic review and meta-analysis on the effects of curative treatments on HRQoL is presented in Chapter 8 and includes 49 studies. A meta-analysis was performed using linear mixed-effects models with quadratic splines to analyse the effects of treatments over time. In the four months after the start of treatment, differences were identified between curative treatment options ($p < 0.001$), with better scores in various HRQoL subscales in patients treated with neoadjuvant chemo(radio)therapy followed by surgery. Over the first year, no differences were identified in HRQoL between curative treatments in both oesophageal and gastric cancer patients. Surgery did not show clinically relevant deterioration of HRQoL in gastric cancer patients, although it did have a clinically relevant impact in oesophageal cancer patients.

In Chapter 9 we report on the development and evaluation of a prediction model for HRQoL in potentially curable patients with oesophagogastric cancer. Prediction models were developed to predict the EORTC summary score at three (N = 537) and six months (N = 480) after the start of treatment. Data was obtained from the prospective POCOP registry (for HRQoL data) and the NCR registry (with patient, tumour and treatment characteristics). The XGBoost algorithm was used to develop the models. Both models had a relatively low performance, with a reduction of only 16% and 11% in RMSE (respectively), compared to a non-informative model. Only baseline summary score was found to be predictive of the summary score at the end of the period in both models. More predictive variables are needed to improve the performance of the models, and further research is needed to determine these variables.

Chapter 10 reports on the development and pilot evaluation study of a web-based tool and training to inform oeosphagogastric cancer patients on treatment outcomes. Treatment outcomes were visualised in a patient-friendly web-interface containing pictographs, barcharts and line graphs. These outcomes are based in part on the prediction models of Chapters 3 and 5, and results from meta-analyses from Chapters 7 and 8. A communication skills training was developed to educate physicians on informing patients on these treatment outcomes. The tool was tested in a focus group, during interviews, and in think-aloud sessions. The tool and training were evaluated in a pilot study, with five physicians and three patients. The preliminary results are promising and show that the tool and training could aid in informing oesophagogastric cancer patients on treatment outcomes in an evidence-based, precise, personalised, and tailored manner.

Concluding remarks

This thesis describes analyses aimed at predicting treatment outcomes in patients with oesophagogastric cancer. Prediction models based on individual patient data are preferred to meta-analyses as they allow for more personalised predictions, yet meta-analyses can provide useful overviews of patterns in large populations. In the future, prediction models can be improved by expanding data sets by including more patients and/or more relevant variables. Establishing the models' quality and limitations is essential. The resulting prediction models and analyses can provide complementary information and visualisations to facilitate shared decision making in clinical practice.

Samenvatting

Voor patiënten met slokdarm- en maagkanker bestaan er verschillende behandelingen die erop gericht zijn om de tumor te verwijderen, klachten te verminderen en de de terugkeer van de tumor te voorkomen. Ondanks deze behandelingen is de overleving van deze patiënten vaak slecht. Ook kan de kwaliteit van leven (KvL) van patiënten verlaagd worden door deze behandelingen, bijvoorbeeld door bijwerkingen en complicaties. Om te bepalen welke behandeling het beste is voor de patiënt en overeenkomt met zijn/haar persoonlijke behandeldoelen, is het cruciaal om de arts en patiënt goed en accuraat te informeren over behandeluitkomsten zoals overleving en KvL. Deze uitkomsten zijn echter moeilijk van tevoren te bepalen en hangen van veel factoren af, zoals patiënt- en tumorkarakteristieken en het type behandeling. Het doel van dit proefschrift is om zo accuraat en objectief mogelijke informatie te geven over behandeluitkomsten voor patiënten met slokdarm- en maagkanker.

Om dit te bereiken, maken we gebruik van twee technieken: meta-analyses en predictiemodellen. Bij meta-analyses worden de data van meerdere studies samengenomen en samengevat om zo een uitspraak te kunnen doen over de uitkomsten van behandelingen (zoals overleving en KvL). Bij een predictiemodel worden gegevens van individuele patiënten gebruikt om een wiskundig model te maken dat voorspellingen kan geven voor (andere) patiënten. In veel gevallen wordt de kwaliteit van een predictiemodel gemeten aan de hand van twee maten: het discriminerend vermogen (c-index) en kalibratie. Bij de c-index wordt gekeken in hoeverre het model in staat is om patiënten van elkaar te onderscheiden die wel of niet een bepaalde behandeluitkomst ondervinden. Deze maat loopt normaal gesproken van 0.5 (geen onderscheidend vermogen) tot 1.0 (perfect onderscheidend vermogen). In het geval dat een overlevingsmodel een c-index van 0.8 heeft, dan betekent dit dat, gemiddeld gesproken, patiënten die langer leven, in 80% van de gevallen een hogere score van het predictiemodel krijgen dan patiënten die minder lang leven. De kalibratie van het model geeft aan in hoeverre de voorspellingen overeenkomen met de werkelijkheid. Bij een perfecte kalibratie hoort een intercept van 0 en een helling van 1.

Hoofdstukken 1 en 11 van dit proefschrift bevatten respectievelijk een algemene inleiding en algemene discussie. Hieronder staat de samenvatting beschreven van de overige hoofdstukken.

In Hoofdstuk 2 geven we een systematische literatuuroverzicht van predictiemodellen die ontwikkeld zijn om behandeluitkomsten te voorspellen voor patiënten met slokdarm- en maagkanker. Er zijn in totaal 47 modellen gevonden. Vijfenveertig studies beschreven de ontwikkeling van deze modellen, en in 16 validatiestudies wordt de kwaliteit van de modellen getoetst. Een meta-analyse van het discriminerend vermogen toont aan dat de c-index van deze modellen redelijk is, met een waarde van 0.75 (variërend van 0.65 tot 0.85). Deze c-index is lager bij validatiestudies dan tijdens de ontwikkeling (0.73 vs. 0.76, $p = 0.01$). De meeste modellen voorspellen overleving na een resectie in patiënten die in opzet curatief behandeld worden, en er zijn geen modellen die KvL voorspellen. Deze studie toont aan dat er een noodzaak is voor modellen die extern gevalideerd zijn, en die zowel de voor- als de nadelen van verschillende behandelingen voorspellen.

In Hoofdstuk 3 beschrijven we de ontwikkeling en validatie van twee nieuwe modellen die de overleving voorspellen in patiënten met uitgezaaide slokdarm- en maagkanker. Twee Cox regressie modellen zijn ontwikkeld op basis van respectievelijk 8,010 slokdarmkankerpatiënten en 4,763 maagkankerpatiënten die gediagnosticeerd zijn tussen 2005 en 2015. Om te bepalen welke variabelen opgenomen worden in het predictiemodel, is er gebruik gemaakt van een expert-panel, ook wel een Delphi consensus genoemd, en het Akaike Information Criterion, dat automatisch op basis van de data een selectie van predictoren maakt. Om de kwaliteit van het model te meten is er gebruik gemaakt van internal-external temporal cross-validation. Bij deze techniek worden er meerdere modellen gemaakt op basis van patiënten die in eerdere jaren gediagnosticeerd zijn, en getoetst worden op patiënten die in latere jaren gediagnosticeerd zijn. Op deze manier kan bepaald worden hoe goed de voorspellingen zijn voor toekomstige patiënten. Hieruit is naar voren gekomen dat de modellen goed gekalibreerd zijn (met een intercept van 0.00 en helling van 1.00). De c-index voor het slokdarmkankermodel was 0.71 (95% CI: 0.70 – 0.71), en bij het maagkankermodel was dit 0.68 (95% CI: 0.67 – 0.69). De modellen hebben redelijke c-indices en kunnen, door het bieden van persoonlijke voorspellingen, gezamenlijke besluitvorming tussen arts en patiënt ondersteunen.

In Hoofdstuk 4 rapporteren we over de externe validatie van de twee modellen die in Hoofdstuk 3 beschreven staan. In deze studie worden de modellen geëvalueerd in een Belgisch cohort van 2,514 slokdarmkanker- en 1,583 maagkankerpatiënten met uitzaaiingen. Het Belgische cohort verschilde van het Nederlandse cohort in meerdere patiënt-, tumor- en behandelkarakteristieken. Het maagkankermodel had in het Belgische cohort een vergelijkbare kalibratie, met een intercept van 0.02 en helling van 0.91, en een vergelijkbare c-index van 0.66. Echter, het slokdarmkankermodel vertoonde een slechtere kalibratie met een intercept van 0.30 en helling van 0.42, en een lagere c-index van 0.64. Deze bevindingen tonen aan dat slokdarmkankermodel niet goed werkt in de Belgische populatie. Het advies is om in toekomstige modellen minder parameters gebruiken om ervoor te zorgen dat de modelprestaties gewaarborgd blijven.

In Hoofdstuk 5 beschrijven we de ontwikkeling en validatie van vier predictiemodellen, gericht op het voorspellen van overleving in slokdarm- en maagkankerpatiënten die uitzaaiingen hebben of in opzet curatief behandeld worden. Deze modellen zijn een update van de modellen beschreven in Hoofdstuk 3, en een uitbreiding van de modellen naar patiënten die in opzet curatief behandeld worden. De Cox regressie modellen zijn gebaseerd op een Nederlands cohort (N = 13,365) met patiënten die tussen 2015 en 2018 gediagnosticeerd zijn. De modellen zijn geëvalueerd door middel van internal-external temporal cross-validation. De c-indices voor patiënten met uitgezaaide kanker zijn respectievelijk 0.72 (95% CI: 0.71–0.74) en 0.73 (95% CI: 0.69–0.75) voor slokdarm- en maagkanker. De modellen voor in opzet curatief behandelde slokdarm- en maagkankerpatiënten hebben een c-index van respectievelijk 0.80 (95% CI: 0.75–0.84) en 0.78 (95% CI: 0.74–0.82). Deze modellen kunnen gebruikt worden voor zowel patiënten met uitgezaaide kanker als in opzet curatief behandelde patiënten, en zijn de eerste modellen voor slokdarm- en maagkanker die overleving voorspellen voor verschillende behandelingen. Er is verder onderzoek nodig om de toegevoegde waarde van het gebruik van deze modellen in de klinische praktijk vast te stellen.

Hoofdstuk 6 beschrijft de ontwikkeling en validatie van een predictiemodel voor overleving in patiënten met een uitgezaaide ductaal adenocarcinoom in de alvleesklier. Op vergelijkbare wijze als in Hoofdstuk 5, is er een Cox regressiemodel ontwikkeld, dat gebaseerd is op een Nederlands cohort van 4,739 patiënten, gediagnosticeerd tussen 2015–2018. Validatie van het model vond plaats door middel van internal-external temporal cross-validation. Het model vertoonde een goede kalibratie en een c-index van 0.72 (95% CI: 0.71-

–0.73). Het model onderschrijft de methodologie die in de voorgaande hoofdstukken beschreven is, en toont aan dat vergelijkbare modelprestaties behaald kunnen worden voor een andere type tumor, wanneer het model op een vergelijkbare manier ontwikkeld en gevalideerd wordt.

In Hoofdstuk 7 rapporteren we een systematische literatuuroverzicht en meta-analyse van KvL in patiënten met slokdarm- en maagkanker die palliatieve systeemtherapie ondergaan. In totaal zijn 43 gerandomiseerde studies geïncludeerd, die de gegevens van in totaal 13,727 patiënten beschrijven. Door middel van lineaire mixed-effects models met kubische splines is er een meta-analyse uitgevoerd op KvL scores. Met deze modellen worden de veranderingen in KvL over tijd gemodelleerd, en wordt het verloop van KvL per behandeling inzichtelijk gemaakt. De analyse toont aan dat patiënten voor aanvang van de behandeling een aangetaste KvL hebben, met een EORTC global health status score van 54.6 (95% CI: 51.9–57.3) in eerstelijnsbehandelingen en een score van 57.9 (95% CI: 55.7–60.1) bij latere behandelingen. Bij de meeste behandelingen bleven de KvL scores constant. In HER2-negatieve tumoren, zorgden combinaties van drie cytotoxische geneesmiddelen (gebaseerd op anthracyclines), of twee cytotoxische geneesmiddelen (gebaseerd op fluoropyrimidine zonder toevoeging van cisplatine), voor betere KvL scores in eerstelijnsbehandelingen in vergelijking met andere eerstelijnsbehandelingen. Bij laterelijnsbehandelingen waren taxanen en doelgerichte behandelingen in staat om te zorgen voor een verbetering in KvL in vergelijking met ondersteunende zorg.

In Hoofdstuk 8 rapporteren we een systematische literatuuroverzicht en meta-analyse gericht op KvL van in opzet curatieve slokdarm- en maagkankerpatiënten. In totaal zijn er 49 studies geïncludeerd. Een meta-analyse van het verloop van KvL over tijd is uitgevoerd door middel van een lineaire mixed-effects models met kwadratische splines. Er werden verschillen gevonden tussen de behandelingen in de eerste vier maanden na de start van de behandeling ($p < 0.001$), en de KvL was bij verschillende subschalen het grootst bij patiënten die neoadjuvante chemo(radio)therapie gevolgd door een resectie ondergingen. Gedurende het eerste jaar na de start van de behandeling werden er geen verschillen gevonden in de KvL van zowel slokdarm- als maagkankerpatiënten. Bij maagkankerpatiënten leidde een operatie niet tot een klinisch relevante verslechtering van Kvl, bij slokdarmkankpatiënten was dit echter wel het geval.

In Hoofdstuk 9 beschrijven we de ontwikkeling en validatie van een predictiemodel om KvL te voorspellen bij in opzet curatief behandelde patiënten met slokdarm- en maagkanker. De modellen voorspelden de EORTC summary score, een maat voor KvL, op drie maanden (N = 537) en zes maanden (N = 480) na de start van behandeling. Voor dit model is gebruikt gemaakt van het prospectieve POCOP register met KvL data, aangevuld met patiënt-, tumor- en behandelkarakteristieken. Het predictiemodel is ontwikkeld met het XGBoost algoritme. Beide modellen presteerden matig in vergelijking met een niet-informatief model, dat de gemiddelde summary score gebruikte als voorspelling. De predictiefout, gemeten met de root mean squared error, nam slechts met respectievelijk 16% en 11% af, in vergelijking met dit niet-informatieve model. Alleen de summary score die gemeten is bij aanvang van de behandeling bleek voorspellend te zijn voor de summary score op drie en zes maanden. Om het model verder te verbeteren zijn meer variabelen nodig die voorspellend zijn voor KvL. Er is verder onderzoek nodig om te bepalen welke variabelen voorspellend zijn voor de summary score.

In Hoofdstuk 10 presenteren we de ontwikkeling en evaluatie van een web-interface en training om patiënten met slokdarm- en maagkanker te informeren over behandeluitkomsten. Deze behandeluitkomsten zijn gepresenteerd in een patiëntvriendelijke web-interface, en worden gevisualiseerd door middel van beelddiagrammen, staafdiagrammen en lijngrafieken. De uitkomsten die gepresenteerd worden zijn deels gebaseerd op de predictiemodelen beschreven in Hoofdstukken 3 en 5, en op de resultaten van de meta-analyses uit Hoofdstukken 7 en 8. Er is ook een communicatietraining ontwikkeld om artsen kennis en vaardigheden bij te brengen over hoe ze patiënten kunnen informeren over behandeluitkomsten. De web-interface is getest door middel van een focus groep, met interviews en tijdens 'hardop denk'-sessies, waarbij gebruikers van de web-interface rechtstreeks feedback geven en hun handelingen toelichten. Ook werden de web-interface en de training geëvalueerd in een pilot studie bij vijf artsen en drie patiënten. De eerste resultaten zijn veelbelovend en tonen aan dat de web-interface en training bij kan dragen aan het informeren van slokdarm- en maagkankerpatiënten over behandeluitkomsten op een objectieve, nauwkeurige, gepersonaliseerde en toegespitste manier.

Slotopmerkingen

Dit proefschrift beschrijft analyses die gericht zijn op het voorspellen van behandeluitkomsten voor patiënten met slokdarm- en maagkanker. Om deze voorspellingen te verrichten, wordt de voorkeur gegeven aan predictiemodellen die op basis van individuele patiëntdata voorspellingen geven, omdat deze meer gepersonaliseerd zijn. Meta-analyses zijn echter ook van toegevoegde waarde, en kunnen patronen in grote populaties beschrijven. De predictiemodellen kunnen in de toekomst verbeterd worden door ze te baseren op grotere datasets. Hierbij kunnen zowel het aantal geïncludeerde patiënten als het aantal relevante predictoren uitgebreid worden. Ook is het essentieel om de kwaliteit en de beperkingen van de modellen vast te stellen. De predictiemodellen en analyses kunnen aanvullende informatie en visualisaties bieden om gezamenlijke besluitvorming tussen arts en patiënt te ondersteunen.

Appendices

List of publications

Publications in this thesis

H.G. van den Boorn, E.G. Engelhardt, J.J. van Kleef, M.A.G. Sprangers, M.G.H. van Oijen, A. Abu-Hanna, A.H. Zwinderman, V.M.H. Coupé and H.W.M. van Laarhoven.
Prediction models for patients with esophageal or gastric cancer: A systematic review and meta-analysis.
PLoS One 13 (2018)

H.G. van den Boorn, A. Abu-Hanna, E. ter Veer, J.J. van Kleef, F. Lordick, M. Stahl, J.A. Ajani, R. Guimbaud, S.H. Park, S.J. Dutton, Y.-J. Bang, N. Boku, N. Haj Mohammad, M.A.G. Sprangers, R.H.A Verhoeven, A.H. Zwinderman, M.G.H. van Oijen and H.W.M. van Laarhoven.
SOURCE: A registry-based prediction model for overall survival in patients with metastatic oesophageal or gastric cancer.
Cancers 11 (2019)

J.J van Kleef*, H.G. van den Boorn*, R.H.A. Verhoeven, K. Vanschoenbeek, A. Abu-Hanna, A.H. Zwinderman, M.A.G. Sprangers, M.G.H. van Oijen, H. de Schutter and H.W.M. van Laarhoven.
External validation of the Dutch SOURCE survival prediction model in Belgian metastatic oesophageal and gastric cancer patients.
Cancers 12 (2020)

J.J. van Kleef, E. ter Veer, H.G. van den Boorn, S. Schokker, L.L. Ngai, M.J. Prins, N. Haj Mohammad, L.V. van de Poll-Franse, A.H. Zwinderman, M.G.H. van Oijen, M.A.G. Sprangers and H.W.M. van Laarhoven.
Quality of life during palliative systemic therapy for esophagogastric cancer: systematic review and meta-analysis.
Journal of the National Cancer Institute 112 (2020)

H.G. van den Boorn*, C.I. Stroes*, A.H. Zwinderman, W.J. Eshuis, M.C.C.M. Hulshof, F.S. van Etten-Jamaludin, M.A.G. Sprangers and H.W.M. van Laarhoven.
Health-related quality of life in curatively-treated patients with oesophageal or gastric cancer: a systematic review and meta-analysis.
Critical Reviews in Oncology/Hematology 154 (2020)

H.G. van den Boorn, A. Abu-Hanna, N. Haj Mohammad, M.C.C.M. Hulshof, S.S. Gisbertz, B.R. Klarenbeek, M. Slingerland, L.V. Beerepoot, T. Rozema, M.A.G. Sprangers, R.H.A. Verhoeven, M.G.H. van Oijen, A.H. Zwinderman and H.W.M. van Laarhoven.
SOURCE: Prediction models for overall survival in patients with metastatic and potentially curable esophageal and gastric cancer.
Journal of the National Comprehensive Cancer Network 19 (2021)

H.G. van den Boorn*, W.P.M. Dijksterhuis*, L.G.M. van der Geest, J. de Vos-Geelen, M.G. Besselink, J.W. Wilmink, M.G.H. van Oijen and H.W.M. van Laarhoven.
SOURCE-PANC: a prediction model for patients with metastatic pancreatic ductal adenocarcinoma based on nationwide population-based data.
Journal of the National Comprehensive Cancer Network 19 (2021)

L.F. van de Water*, H.G. van den Boorn*, F. Hoxha, I. Henselmans, M.M. Calff, M.A.G. Sprangers, A. Abu-Hanna, E.M.A. Smets and Hanneke W.M. van Laarhoven.
Informing patients with esophagogastric cancer about treatment outcomes by using a web-based tool and training: development and evaluation study.
Journal of Medical Internet Research 23 (2021)

H.G. van den Boorn, S.C. Kuijper, J.J. van Kleef, R.H.A. Verhoeven, S.S. Gisbertz, W. Jansen, J.P. Ruurda, C. Rosman, M. Slingerland, A.H. Zwinderman, M.A.G. Sprangers, H.W.M. van Laarhoven and A. Abu-Hanna.
Prediction of health-related quality of life in patients with potentially curable oesophageal or gastric cancer.
Submitted

Other peer reviewed publications

L.L. Ngai, E. ter Veer, H.G. van den Boorn, E. Hugo van Herk, J.J. van Kleef, M.G.H. van Oijen and H.W.M. van Laarhoven.
TOXview: a novel graphical presentation of cancer treatment toxicity profiles.
Acta Oncologica 58 (2019)

T. van den Ende*, F.A. Abe Nijenhuis*, H.G. van den Boorn, E. ter Veer, M.C.C.M. Hulshof, S.S. Gisbertz, M.G.H. van Oijen and H.W.M. van Laarhoven.
COMplot, a graphical presentation of complication profiles and adverse effects for the curative treatment of gastric cancer: a systematic review and meta-analysis.
Frontiers in Oncology 9 (2019)

L.F. van de Water, J.J. van Kleef, W.P.M. Dijksterhuis, I. Henselmans, H.G. van den Boorn, N.M. Vaarzon Morel, K.F. Schut, J.G. Daams, E.M.A. Smets and H.W.M. van Laarhoven.
Communicating treatment risks and benefits to cancer patients: a systematic review of communication methods.
Quality of Life Research 29 (2020)

T. van den Ende, H.G. van den Boorn, N.M. Hoonhout, F.S. van Etten-Jamaludin, S.L. Meijer, S. Derks, T.D. de Gruijl, M.F. Bijlsma, M.G.H van Oijen and H.W.M. van Laarhoven.
Priming the tumor immune microenvironment with chemo(radio)therapy: a systematic review across tumor types.
Biochimica et Biophysica Acta - Reviews on Cancer 1874 (2020)

J.J. van Kleef, W.P.M. Dijksterhuis, H.G. van den Boorn, M. Prins, R.H.A. Verhoeven, S.S. Gisbertz, M. Slingerland, N. Haj Mohammad, G.-J. Creemers, K.J. Neelis, J. Heisterkamp, C. Rosman, J.P. Ruurda, E.A. Kouwenhoven, L.V. van de Poll-Franse, M.G.H. van Oijen, M.A.G. Sprangers and H.W.M. van Laarhoven.
Prognostic value of patient-reported quality of life for survival in oesophagogastric cancer: analysis from the population-based POCOP study.
Gastric Cancer (2021)

C. Jongerius, H.G. van den Boorn T. Callemein, N.T. Boeske, J.A. Romijn, E.M.A. Smets and M.A. Hillen.
Eye-tracking analyses of physician face gaze patterns in consultations.
Scientific Reports 11 (2021)

* *These authors contributed equally*

PhD portfolio

Name PhD student:	Héctor G. van den Boorn
PhD period:	01-01-2017 – 31-12-2020
PhD supervisor:	Prof. dr. H.W.M. van Laarhoven

	Year	Workload (ECTS)
1. PhD training		
General courses		
- Practical biostatistics	2017	1.4
- Searching for a systematic review	2017	0.1
- World of science	2017	0.7
- Bedrijfshulpverlening	2017	1
- Advanced topics in biostatistics	2018	2.1
Specific courses		
- Introduction to cost-effectiveness analysis (Society for Medical Decision Making)	2018	0.2
- Basiscursus oncologie (Nederlandse Vereniging voor Oncologie)	2018	2.2
- Intro to patient-reported outcomes (International Society for Quality of Life Research)	2018	0.25
Oral presentations		
- "A novel prediction model for overall survival in patients with metastatic esophageal or gastric cancer". Cancer Center Amsterdam Annual Retreat, Noordwijkerhout, The Netherlands.	2018	1
- "Design and evaluation of a user-friendly web-interface with prediction models". Cancer Center Amsterdam Annual Retreat, Noordwijkerhout, The Netherlands.	2019	1
Poster presentations		
- "A machine learning approach to predict the outcome of intervention programs for children with unilateral cerebral palsy". European Academy of Childhood Disability Annual Conference, Amsterdam.	2017	0.25
- "Two novel registry-based prediction models for overall survival in patients with metastatic esophageal or gastric cancer". Society for Medical Decision Making Annual Conference, Leiden, The Netherlands.	2018	0.75
- "Designing a user-friendly web-interface with prediction models for survival, health-related quality-of-life and toxicity for cancer patients". International Society for Quality of Life Research Annual Conference, Dublin, Ireland.	2018	1
- "Prediction models for overall survival in patients with metastatic and potentially curable esophageal and gastric cancer (the SOURCE study)". American Society of Clinical Oncology GI Conference, San Francisco, United States of America.	2020	0.75
2. Teaching		
Lecturing		
Fundamentals of data science in medicine (pracatical supervision)	2018–2020	3
Supervising		
Florian Hoxha, student scientific internship,"Visualizing predictions of survival, quality of life and toxicity for cancer patients and oncologists with a web-application" (department of medical informatics, AMC)	2018	2

About the author

Héctor van den Boorn was born in 1991 on Curaçao, The Netherlands Antilles. In 1992 his family moved to The Netherlands, where he grew up in the city of Hoorn. He attended his secondary education at Oscar Romero in Hoorn, and graduated cum laude from Gymnasium in 2009. He then moved to Nijmegen, where he studied at the Radboud University Nijmegen. During his studies, he obtained his Bachelor's degree in Artificial Intelligence cum laude in 2013. He attended the Interdisciplinary Radboud Honours Academy from 2010-2012, and in 2012 he studied Computer Science for one semester at San Diego State University. In 2016, he obtained his Master's degree in Artificial Intelligence cum laude with a specialisation in 'Computation in neural and artificial systems'. He wrote his Master thesis on the use of machine learning for the prediction of treatment benefit in children with cerebral palsy.
During his studies, he worked as a research assistant at the Max Planck Institute for Psycholinguistics and the Donders Institute for Brain, Cognition and Behaviour in Nijmegen. Additionally, he worked as a teaching assistant at the Radboud University Nijmegen. In 2017, he obtained the position of PhD student at the Academic Medical Centre in Amsterdam, conducting research on the prediction of treatment outcomes in oesophagogastric cancer patients, described in this thesis. Currently, he works as a data scientist at the Tax and Customs Administration of The Netherlands.