Reusing routine cancer care data for registries and decision support

Melle Sjoerd Sieswerda

Reusing routine cancer care data for registries and decision support

Melle Sjoerd Sieswerda

Provided by thesis specialist Ridderprint, ridderprint.nl
Printing: Ridderprint
Layout and design: Henry Smaal, persoonlijkproefschrift.nl

Reusing routine cancer care data for registries and decision support

DISSERTATION
to obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus, Prof. Dr. Pamela Habibović,
in accordance with the decision of the Board of Deans,
to be defended in public on Friday, April 4th 2025, at 13:00 hours

by Melle Sjoerd Sieswerda

Supervisors:
Prof. Dr.ir. A.L.A.J. Dekker – Maastricht University
Prof. Dr. V.E.P.P. Lemmens – Maastricht University

Co-supervisor:
Dr. I. Bermejo – Hasselt University

Assessment Committee
Chair:
Prof. Dr. D.K.M. De Ruysscher – Maastricht University

Members:
Dr. S.O. Breukink - Maastricht UMC+
Prof. Dr. P.J.F. Lucas - University of Twente
Prof. Dr. R. Verheij - Tilburg University/NIVEL

Organizations that contributed to the doctoral research:
Netherlands Comprehensive Cancer Organisation (IKNL)

“Felix, qui potuit rerum cognoscere causas”
-- Vergilius, Georgica (ca. 29 v.Chr.)

Index

1 Introduction

General introduction

Epidemiology of cancer

In 2019 over 120.000 people were newly diagnosed with cancer in the Netherlands.[1] Over 45.000 people died of cancer in the same year, making it the primary cause of death (with cardiovascular disease being a close second).[2] As such, improving cancer diagnosis and treatment will have significant impact on population health.

Pathophysiology of cancer

Normally, cell replication is a carefully regulated process: cells only replicate when needed, perform self-checks before doing so, and even self-destruct if their integrity is compromised beyond repair (a process known as apoptosis). They also communicate with each other: feedback processes exist to stop tissue growth when no longer needed.

In cancer these fail-safes are broken, and cells replicate unchecked. The root cause lies in genetic mutations, which may cause cellular functions that inhibit replication to be lost, or functions that promote replication to be gained. Another function that may be gained is the ability to invade adjacent tissues or for cells to detach from the primary tumor and settle someplace else in the body, leading to metastases.[3]

Which genes are actively used by a cell ("expressed") depends on tissue type and cellular function (among other factors). As a result, susceptibility to specific mutations varies between cells as well: a mutation in a gene that is inactive, will have little effect. Genetic mutations can be hereditary or acquired. When acquired, they can be the result of a natural process of cell division, but environmental and lifestyle factors, like alcohol or smoking, are increasingly important.

Treatment

When it comes to cancer treatment, three modalities can be discerned: surgery, chemotherapy, and radiotherapy (which includes proton therapy). While exact treatment (and prognosis) depends on both tumor type and stage, surgery is the cornerstone of curative treatment. This is emphasized by the fact that when chemotherapy or radiotherapy is combined with surgery, they are considered adjuvant: neoadjuvant when given preceding surgery or (just) adjuvant when given afterwards. Neoadjuvant therapy is frequently given with the aim to reduce tumor size prior to surgery, improving the possibility of a complete resection (or even making surgery possible). Adjuvant therapy generally targets undetectable (micro)metastases that might remain after all detectable disease has been (surgically) removed.

The Netherlands Cancer Registry

Given the impact of cancer on national health, many countries monitor incidence, treatment, and outcomes. For this purpose, the Netherlands Comprehensive Cancer Organization (IKNL) maintains the Netherlands Cancer Registry (NCR). This national, population-based database contains data collected from (electronic) healthcare records on diagnosis, tumor characteristics, initial treatment, and outcome.

The NCR covers over 95% of all cancer occurrences in the country, goes back over 30 years, and contained over 2.5 million patient records in 2019[1,a]. This makes it a valuable resource for (clinical) research, quality monitoring, benchmarking, and policy decisions.

Extracting routine clinical data for re-use in research

Evidently, every (statistical) analysis, every attempt to monitor quality, and every benchmark requires data. While historically, this data is manually collected[b], technological advances have made their mark here as well. Since 2016, all hospitals in the Netherlands use Electronic Healthcare Records (EHRs). This holds the promise to unlock health data, routinely collected in clinical use, for secondary purposes.

Currently, IKNL employs trained registrars that extract data from hospital medical records to populate the NCR. When registration is limited to a (relatively) small set of data points, this works very well. However, with the increase in available processing power and storage capacity, the demand for additional variables has increased. Clearly this process is hard (or at least: costly) to scale up. As such, it seems desirable to automate the registration process.

A similar need is felt by hospitals that must submit data to quality registrations. According to a recent survey, hospitals in the Netherlands participate in 60 quality registrations on average.[c] Interestingly, tumor boards present a unique opportunity to extract data from routine care. These are meetings where medical specialists from different disciplines (e.g., radiology, pathology, surgery, radiotherapy, etc.) discuss diagnostic findings and decide on treatment options. This decision-making requires aggregation and consolidation of clinical data, which, when well documented, forms a valuable source of information for research. Unfortunately, most tumor boards report using free text and quality of the clinical documentation varies wildly.

a The estimate for 2023 is over 3 million, but incidence registration is in progress at the time of writing.
b The Iris dataset, collected by Fisher and Anderson in the 1930s, which contains data on morphologic features of three types of Iris flowers, is an example that is still used today.[4,5]
c https://www.atr-regeldruk.nl/wp-content/uploads/2023/06/19-U052-Min-VWS-Uitvoeringsregeling-Wkkgz-ivm-regie-op-kwaliteitsregistraties-w.g.pdf

In Chapter 2 we explored the possibility of increasing the quality of clinical documentation and facilitate automatically populating the NCR, by standardizing tumor board reports according to the Dutch national clinical practice guideline for breast cancer. Since introducing structured reporting runs the risk of adding to the clinical workload, we additionally investigated the impact on the clinical workflow, and time spent on tumor board preparation and clinical documentation in breast cancer.

Federated Learning: using (clinical) data across borders

Unfortunately, in many situations even a nationwide, population-based registry does not contain enough patients. This may feel counterintuitive, given that the NCR alone contains over 2.5 million patients, but is easily understood when one realizes that cohort selection can quickly reduce the dataset. For example, of these 2.5 million, only ~13.000 patients had liver cancer, of which only ~4.000 were diagnosed after 2012 of which only ~500 were women ≤ 65yo.

Another reason why data from a single country can be insufficient, is because of lack of variability. Clinical Practice Guidelines (CPGs) are increasingly used to drive treatment decisions. While this is done with the aim of increasing health care quality by reducing *unwanted* variability, a side effect is that treatment groups become more homogeneous.

Internationally, a lot of data is available that may complement the NCR. For example, the database from the Surveillance, Epidemiology, and End Results (SEER) Program. However, record level data cannot easily be shared across borders, due to legal and privacy constraints.

Federated Learning is a technique that solves these issues by enabling statistical analyses without sharing record level data; the data remains at the source, and only aggregated results and statistics are exchanged.[6–8] Interestingly, several algorithm classes, such as the generalized linear models, have been shown to be decomposable in such a way that the result is mathematically equivalent to their regular implementation.[9] So, while the approach is much like a meta-analysis, the result is as if all data had been pooled and centrally analyzed.

To apply federated learning, two things are required: 1) an infrastructure that facilitates communication between the participants, and 2) algorithms that use the infrastructure to perform (statistical) analysis. In Chapter 4 we describe how we implemented a flexible, programming language agnostic, infrastructure for federated learning[d], as well as the algorithm that calculates the Cox Proportional Hazards model. Both are employed to investigate incidence and treatment in oral cavity cancer in the Netherlands and Taiwan.

d This software would later become vantage6.

A quick overview of recent developments

Over the past years, propelling new concepts like federated learning, there have been substantial developments in computer science. The increase of computational power, storage capacity, and data availability have enabled algorithms that were previously intractable.

These developments have inspired the field of machine learning: an area of research that focusses on computer algorithms that imitate the way that people learn from experience: by trial and error. Usually, this involves the minimization of an error function (that describes the goodness of fit of the model) using an iterative process.

These days, many algorithms or algorithm classes exist, such as Support Vector Machines, Random Forests, XGBoost, Neural Networks, and Bayesian Networks.[10–15] At the time of writing, a specific neural network, the large language model GPT-4, has made international headlines with its ability to generate text.[16] Further testing and development are required, but these networks already seem capable of answering questions, as well as rephrasing and summarizing text.

While the impact that neural networks like these will have on society can hardly be overestimated, they also have a downside: these models are so complex that we cannot inspect their parameters and verify their inner workings. At the same time, especially generative multipurpose networks like GPT-4, are known to "hallucinate" and produce results that *seem* plausible but are pure fiction.[16] Additionally, or as a result, they are also not well suited for causal analysis. Neural networks are essentially black boxes. For many applications this is fine, but in some areas of research, such as the medical domain, this poses a challenge. In these areas Bayesian Networks, which are further introduced in the following section, may provide an alternative that is easy to understand and opens the door to causal analysis.

A brief introduction to probability theory, conditional (in)dependence, and Bayesian Networks

Bayesian Networks (BNs) are probabilistic graphical models, based on the idea that cause and effect are not always absolute (deterministic). To understand the basic concepts behind BNs, it is useful to first introduce Bayes' theorem of conditional probability and conditional (in)dependence. Since BNs are mostly used with discrete random variables, we will limit the introduction to this use case.

Probability and Joint Probability

In statistics, events are modeled with random variables: variables with an accompanying probability distribution. In the discrete case, this probability distribution consists of a table that describes the probability of each state. For example, when considering a student's grade (measured as A, B, or C) for a class, we might find probabilities of 25%, 37%, and 38% for an A, B, or C respectively (Table 1).[e]

Of course, there may be variables that influence these probabilities, such as a student's intelligence, or the class's difficulty. If we stratify the distribution over G by intelligence, denoted here as "high" or "low", we obtain the Joint Probability distribution over I and G (Table 2).

The probability of a student obtaining an A is denoted as $P(G = g_1) = 0.25$. This can be shortened to $P(g_1) = 0.25$ if it is clear that g_1 belongs to distribution G. The full distribution is written as $P(G)$. Similarly, the joint distribution over I and G is denoted as $P(I, G)$.

$P(G)$	
g_1 (A)	0.25
g_2 (B)	0.37
g_3 (C)	0.38
	1.0

Table 1: Probability distribution $P(G)$ for the variable Grade.

$P(I, G)$		I (intelligence)		$P(G)$
		i_0 (low)	i_1 (high)	
G (grade)	g_1 (A)	0.07	0.18	0.25
	g_2 (B)	0.28	0.09	0.37
	g_3 (C)	0.35	0.03	0.38
$P(I)$		0.70	0.30	1.0

Table 2: Joint Probability distribution of $P(I, G)$ for variables Grade and Intelligence. The rightmost column and bottom row display the marginal probabilities, that is: the probabilities for Grade and Intelligence separately.

e Example copied from Koller and Friedman's Probabilistic Graphical Models.[13]

Conditional probability

Using a joint probability table (JPT) we can look up the probability that a student is both intelligent *and* obtained a good grade ($P(g_1,\ i_1) = 0.18$). We can use this table to calculate the probability that an intelligent student obtained a good grade, the conditional probability of g_1 *given* i_1. Mathematically this is denoted with the conditioning variables written behind a vertical dash, for example $P(g_1 \mid i_1)$ or $P(G \mid I)$ for the full distribution.

The formula for calculating conditional probabilities is called Bayes' theorem (Equation 1), named after Reverend Thomas Bayes (~1701-1761).[17] The theorem states that the probability of event A given B is equal to the joint probability of A and B, divided by the probability of B. Following the student example, this would equate to $P(g_1 \mid i_1) = P(g_1,\ i_1)/P(i_1) = 0.6$.

$$P(A \mid B) = \frac{P(A, B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

Equation 1: Bayes' theorem. $P(A \mid B)$ denotes the conditional probability distribution of A given B. $P(A, B)$ denotes the joint probability distribution of A and B. $P(B)$ represents the (prior) probability distribution of B.

In the case of intelligence and grade, it is easy to see a causal relationship, where intelligence drives grade. From an information point of view, however, the reverse relationship is equally relevant: knowing a student obtained a good grade increases the probability the student has high intelligence. This is reflected in Bayes' formula: calculating the reverse probability $P(B \mid A)$ merely entails a division by $P(A)$ and multiplication by $P(B)$.

Dependence and independence in statistics

When knowledge about variable A does not convey any information about variable B, these variables are said to be statistically independent. This is denoted as $A \perp\!\!\!\perp B$. For example, when rolling two dice, A and B, knowing the value of one will not yield any information about the value of the other. In this case $P(A \mid B) = P(A)$ and vice versa. The opposite of independence is, of course, statistical dependence which is denoted as $A \not\perp\!\!\!\perp B$.

Both dependence and independence can be conditional too. Continuing the previous example, assume we have three variables that model the relationship between the difficulty of a university course (D), the grade that a student receives for the course (G), and whether the professor is willing to write a letter of recommendation for the student (L). Additionally, it seems reasonable to assume the student's grade G will depend on the course difficulty D, and that the teacher's willingness to write a letter L only depends on the grade. As a result, D and L are (indirectly) associated and thus $D \not\perp\!\!\!\perp L$. However, if we know the student's grade, knowing the course difficulty will not yield any information about the likelihood of the teacher's willingness to write the letter. In other words, if G is known, the association between D and L disappears, so D is conditionally independent of L given G. This is denoted as $D \perp\!\!\!\perp L \mid G$.

Bayesian Networks

Bayesian Networks (BNs) combine the concepts of conditional probability and (in) dependence with a visual representation that uses directed acyclic graphs (DAGs): nodes represent variables and edges between the nodes are used to express the relationships between them. Each node is associated with a probability distribution that is dependent on its parents. Continuing the example above, node D would be associated with a probability distribution $P(D)$, G with $P(G \mid D)$, and L with $P(L \mid G)$. These three (conditional) distributions together define the joint probability distribution $P(D,\ G,\ L)$.

Figure 1 shows how these relationships would be visualized in a BN, adding the variables: I (intelligence) and S (SAT-score). Intelligence (I) influences both the grade (G) the student might obtain for the course and the SAT-score (S), which reflects earlier (scholarly) performance.

In a BN it is possible to set evidence on one or more of the nodes. This triggers an update and calculates the conditional probabilities of the remaining nodes *given* the evidence. For example, if we input $I = i_1$, the BN would calculate $P(D \mid i_1)$, $P(G \mid i_1)$, $P\left(S \middle| i_1\right)$, and $P(L \mid i_1)$. Interestingly, every node can be used as input or output.

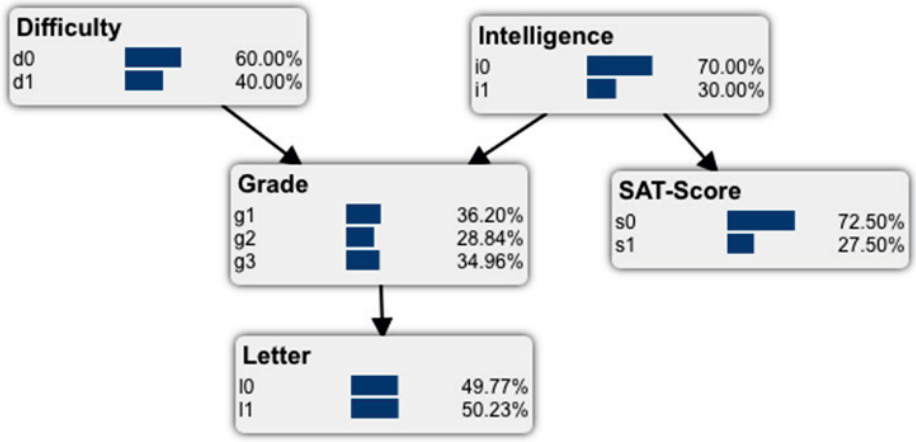


Figure 1: the Student Network (copied from Koller and Friedman). Each node represents a variable. The edges denote associations between variables, pointing from parents towards children. Each node is associated with a probability distribution (conditional probability table, or CPT) that is conditional on its parents. These CPTs can be used to compute the prior probabilities for each node as shown here.

Building blocks of Bayesian Networks

Looking at this graph, three building blocks, each consisting of three variables, can be identified: the chain, common cause, and common effect. Interestingly, these building blocks are directly related to (conditional) independencies.

Chain, or mediator

In a chain one variable influences another through an intermediary. This intermediary is known as a mediator. For example, G would be a mediator in the chain $D \rightarrow G \rightarrow L$ from Figure 1. In a chain, the first and last variables are conditionally independent given the mediator.

Common cause, or confounder

If two variables are influenced by the same variable, they share a common cause. For example, in Figure 1 the variable I is a parent of both G and S and acts as a common cause: $G \leftarrow I \rightarrow S$. In this situation the first and last variables are conditionally independent given the common cause. Common causes are also known as confounders (see also the section "What is confounding?" below).

Common effect, or collider

If two parents both influence another variable, they are said to have a common effect. The variable that is affected, is known as a collider. For example, in Figure 1 the variable G has parents D and I: $D \rightarrow G \leftarrow I$. This building block is also known as a V-structure, which lends its name from the V-shape that appears if a graph is depicted with the parents on top and the children below.

In terms of dependency this configuration differs from the chain and common cause: without prior information, the first and last variable are *in*dependent, but become conditionally dependent given the collider.

Creating Bayesian Networks

Creating BNs can be broken down into two steps: 1) definition of the structure (i.e., the graph), and 2) determining the parameters of the associated probability distributions. While it is possible to manually define both structure and parameters, for example through guidance from field experts, it is also possible to derive them from data using algorithms. If data is used, the first step is referred to as structure learning and the second step is known as parameter estimation. Hybrid options, such as manually defining structure learning constraints or defining the prior probabilities for parameter estimation, are also possible.

Dealing with changes in classification systems (over time)

Hybrid models, as described above, are especially interesting in situations where data availability is limited, but human knowledge about the underlying data model is at hand. For example, when a classification system is revised, there is likely some understanding of how the new version compares to the previous, and which categories have been split or merged. This may be helpful in developing a model that "translates" between versions when using multiple versions in an analysis is not feasible or desirable. In oncology, the TNM classification system provides such a use case.

The TNM classification system

Both quantitative and qualitative analytics with respect to cancer require clear definitions of disease stage. To facilitate this, the TNM system for classification of malignant (solid) tumors was developed.[18, 19] It not only aids in stratifying patients for research, but also helps clinicians assess prognosis and guides treatment decisions.

The TNM system works by classifying the disease along three axes: characteristics and extent of the tumor (T descriptor), involvement of local lymph nodes (N descriptor), and the presence (and location) of distant metastases (M descriptor). For example, a patient with a tumor in the colon limited to the submucosa (T_1), without lymph node involvement (N_0), and without distant metastases (M_0) would be classified as $T_1N_0M_0$. The possible values each descriptor can take differ per tumor type: the categories for colon cancer are different from those for lung cancer.

The total number of combinations of T, N, and M values can quickly become impractical. For example, for colon cancer there would be 108 (6 x 6 x 3) possible combinations, some of which are very unlikely (e.g., $T_0N_0M_1$). To keep things manageable, rules have been created that define stage groups, denoted by roman numerals ranging from I to IV; subdivisions are denoted using a character suffix. Again, these definitions are tumor specific. Generally, prognosis is worse for increasing stages (i.e., stage I has a better prognosis than stage III), and stage IV is associated with metastasized disease.

Revisions of the TNM classification system

Of course, our understanding of the pathophysiology, diagnosis, and treatment of cancer continuously evolves. Relationships between variables are better understood and new variables are identified. This inherently means that classification systems, which are based in current knowledge, need the ability to change too.

For this reason, the TNM system is revised every 5.7 years on average. During revision, which includes both the individual descriptors (T, N, and M) and the stage groupings, changes may involve (combinations of) redefinition, introduction, removal, splitting, or merging of classes.[20] As a result, classes with the same label are not necessarily equivalent across editions.

These differences between TNM editions (versions) can complicate analyses across edition-boundaries. To tackle this issue, conceptually two solutions are available: 1) mapping and 2) full reclassification. The first solution requires some form of compatibility between two editions: it should be possible to define rules that "translate" edition A into edition B, which is easy if two groups in the source edition are merged into one in the target edition. Reclassification requires additional data (variables) to classify the patient in the target edition. Of course, combinations of mapping and reclassification are possible too. Unfortunately, mapping and/or reclassification is not always possible. Most likely because these additional variables are just not available.

Probabilistic reclassification with Bayesian Networks

In oncology, analyses frequently span a decade: using 10-year survival as an outcome is not uncommon. As soon as TNM-stage is involved, this likely requires dealing with 2-3 editions of the classification system, which complicates analysis.

BNs can help alleviate this issue by enabling probabilistic reclassification through a hybrid approach: combining expert knowledge with real-world data. In Chapter 4 we describe how the prognostic capabilities of the TNM classification even make it possible to treat the problem of training the classifier as a latent class or clustering analysis.

Another area where BNs have interesting uses, is in estimating treatment effect in observational data through causal analysis. To see how we can go from correlation to causation requires a deeper understanding of confounding, which is further explained in the next section.

Identifying and mitigating confounders to estimate causal (treatment) effects

What is confounding?

In statistics the mantra "correlation is not causation" is repeated often, and (mostly) rightly so. Still, even without considering causation, correlation can help us predict outcomes from observational data, if we stick to *observing* only. For example, if ice-cream sales were high yesterday, we might predict a simultaneous increase in number of sunburns.

However, if we try to estimate the effect of handing out free ice-cream (an intervention) on the number of sunburns, the previously found correlation is useless because it was confounded by the weather. There is no direct causal effect between sales and sunburns, but there exists a *common* cause to both outcomes; handing out ice-cream will not affect the weather.

Essentially any variable with a causal effect on both the intervention, such as treatment selection, and the outcome of interest can be considered a confounder. The simplest example of confounding involves three variables X (intervention), Y (outcome), and Z (confounder), where Z influences both X and Y ($Z \rightarrow X$, $Z \rightarrow Y$; Figure 2).

Confounding explains why it is (relatively) easy to use prediction models to estimate survival for patients that have already decided on treatment, but more difficult to predict treatment effect. For instance, assume we are doctors and would like to compare the effect of drugs A and B using historical data. Assume further that, historically, we considered relatively young and healthy patients fit enough to take drug A, but those in poor condition were thought more suitable for drug B. If we then (naïvely) try to estimate treatment effect on overall survival, we might see that patients on drug A perform much better. However, they had a better prognosis to begin with: any measured treatment effect of drug A will be skewed toward better outcomes. We may not be measuring treatment effect, but rather prior health status. Clearly, this is undesirable.

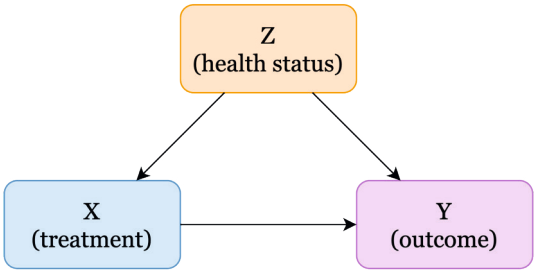


Figure 2: Example of confounding. Choice for treatment (X) is influenced by prior health status (Z), which also (partially) determines outcome (Y).

Mitigating confounding

When conducting a prospective, controlled experiment, confounding can be avoided by using randomization: if treatment is determined by random allocation, this breaks the link with prior health status. In observational data, where treatment is not randomly allocated, correcting for confounders would achieve the same thing, and is conceptually straightforward if (all) confounders are known. Solutions include adding confounding variables to regression formulas, stratification (e.g., calculating treatment effect for each age group), and propensity score matching. For BNs even a formal methodology for causal reasoning exists: the do-calculus.

Identifying confounders

In practice, the difficulty in correcting for confounding has two reasons: 1) not all confounders may have been measured, and 2) there may be uncertainty about which measured variables actually act as confounders. The first reason is a fundamental, but situational issue. Depending on their level of understanding of a domain, a researcher

may feel either more or less certain that all relevant variables are accounted for or that an important variable remains unmeasured. The second issue may be mitigated by using one of several criteria for identification of confounders, for example the pre-treatment criterion, common-cause criterion, backdoor-path criterion, or disjunctive cause criterion.[21, 22] The pre-treatment criterion would select all pre-treatment variables as confounders. The common-cause criterion would only correct for those variables thought to be common causes of exposure (treatment selection) and outcome. The backdoor-path criterion resembles the common-cause criterion but takes chains of influence into account and can thus correct for indirect confounders. The disjunctive cause criterion would correct for those variables thought to be either a cause of exposure or outcome.

Correcting for the wrong set of potential confounders is not without risk. It may lead to dilution of statistical power or even introduce bias.[21, 22] When there is uncertainty about the presence of unmeasured confounding, it can be reasoned that, generally, the disjunctive cause or backdoor-path criteria yield the most unbiased results.

Unfortunately, application of these criteria requires a causal model, which is usually not available, and deriving a causal model from data alone is not straightforward. Especially determining direction of influence is problematic and needs additional information. Manual definition of a causal model would be an option, if it were not that associations between variables are frequently a point of contention. For example, an association between age and treatment may be *suspected*, but is not necessarily a given. When multiple variables are involved, describing all their relationships quickly becomes complex (and uncertain).

Interestingly, the ability of BNs to incorporate expert knowledge provides a hybrid solution for causal model development. Not necessarily by enforcing (directed) edges, but by precluding directed edges that are *known* to be non-causal. For example, cause will always precede effect: *treatment* can never cause *age*, and that directed edge can be precluded. For many other variables, similar arguments can be made. When combining this prior knowledge, in the form of a blacklist, with structure learning algorithms for BNs, it is possible to obtain data-driven causal models. This is illustrated in Chapters 5 and 6, where we apply structure learning algorithms for BNs to obtain causal models for prostate and colon cancer respectively.

Guide to reading this thesis

In **Chapter 2** we explored the possibility of standardizing tumor board reports according to the Dutch national clinical practice guideline for breast cancer. The goal was to increase the quality of the clinical documentation and enable secondary use for research.

Subsequently, in **Chapter 3**, we show how BNs can be used to facilitate probabilistic reclassification, enabling statistical analyses across TNM-editions in lung cancer.

To further increase the available data by applying federated learning, two things are required: 1) an infrastructure that facilitates communication between the participants, and 2) algorithms that use the infrastructure to perform (statistical) analysis. In **Chapter 4** we describe how we implemented a flexible, programming language agnostic, infrastructure for federated learning, as well as the algorithm that calculates the Cox Proportional Hazards model. Both are employed to investigate incidence and treatment in oral cavity cancer in the Netherlands and Taiwan.

Then **Chapters 5, 6, and 7** investigate the possibility of using structure learning algorithms to learn from cancer registry data, causal Bayesian Networks for prostate cancer, colon cancer, and rectal cancer respectively. The end-goal being reliable estimation of treatment effect using observational data.

Finally, in **Chapter 8** we will revisit the major outcomes, and discuss limitations and implications of our work.

References

1. NCR Data | Incidence - Graph [Internet][cited 2023 Aug 23] Available from: https://nkr-cijfers.iknl.nl/#/viewer/3bda9d10-71dc-4f15-830f-286f590d1b39
2. NCR Data | Mortality - Graph [Internet][cited 2023 Aug 23] Available from: https://nkr-cijfers.iknl.nl/#/viewer/be54d472-54e7-4fd2-b26f-35f7f8273118
3. Hanahan D, Weinberg RA: The Hallmarks of Cancer. Cell 100:57–70, 2000
4. Anderson E: The Species Problem in Iris. Ann Mo Bot Gard 23:457–509, 1936
5. Fisher RA: The Use of Multiple Measurements in Taxonomic Problems. Ann Eugen 7:179–188, 1936
6. Lu C-L, Wang S, Ji Z, et al: WebDISCO: a web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc JAMIA 22:1212–1219, 2015
7. Wu Y, Jiang X, Kim J, et al: Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. J Am Med Inform Assoc JAMIA 19:758–764, 2012
8. Jiang W, Li P, Wang S, et al: WebGLORE: a web service for Grid LOgistic REgression. Bioinforma Oxf Engl 29:3238–3240, 2013
9. Cellamare M, van Gestel AJ, Alradhi H, et al: A Federated Generalized Linear Model for Privacy-Preserving Analysis. Algorithms 15:243, 2022
10. Chen T, Guestrin C: XGBoost: A Scalable Tree Boosting System [Internet], in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016, pp 785–794[cited 2024 Feb 1] Available from: http://arxiv.org/abs/1603.02754
11. Cortes C, Vapnik V: Support-vector networks. Mach Learn 20:273–297, 1995
12. Pearl J: Causality: Models, Reasoning and Inference (ed 2nd). New York, NY, USA, Cambridge University Press, 2009
13. Koller D, Friedman N: Probabilistic Graphical Models: Principles and Techniques. Cambridge, Massachusets; London, England, MIT Press, 2009
14. 30 years of adaptive neural networks: perceptron, Madaline, and backpropagation | IEEE Journals & Magazine | IEEE Xplore [Internet][cited 2024 Feb 2] Available from: https://ieeexplore.ieee.org/document/58323
15. Bishop CM: Neural Networks for Pattern Recognition. USA, Oxford University Press, Inc., 1995
16. OpenAI, Achiam J, Adler S, et al: GPT-4 Technical Report [Internet], 2023[cited 2024 Feb 2] Available from: http://arxiv.org/abs/2303.08774
17. Bellhouse DR: The Reverend Thomas Bayes, FRS: A Biography to Celebrate the Tercentenary of His Birth. Stat Sci 19:3–32, 2004
18. Edge SB, American Joint Committee on Cancer (eds): AJCC Cancer Staging Manual (ed 7). New York, Springer, 2010
19. Amin MB, Edge S, Greene F, et al (eds): AJCC Cancer Staging Manual (ed 8). Springer International Publishing, 2017
20. Goldstraw P, Crowley J, Chansky K, et al: The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours. J Thorac Oncol Off Publ Int Assoc Study Lung Cancer 2:706–714, 2007
21. VanderWeele TJ, Shpitser I: A new criterion for confounder selection. Biometrics 67:1406–1413, 2011
22. Häggström J: Data-driven confounder selection via Markov and Bayesian networks. Biometrics 74:389–398, 2018

2

Impact on Quality of Documentation and Workload of the Introduction of a National Information Standard for Tumor Board Reporting

Adapted from Ebben K.C.W.J., Sieswerda M.S., Luiten E.J.T., Heijns J.B., van der Pol C.C., Bessems M., Honkoop A.H., Hendriks M.P., Verloop J., Verbeek X.A.A.M.; "**Impact on Quality of Documentation and Workload of the Introduction of a National Information Standard for Tumor Board Reporting**"; *JCO Clin Cancer Inform*. 2020;4:346-356.

Abstract

PURPOSE Tumor boards, clinical practice guidelines, and cancer registries are intertwined cancer care quality instruments. Standardized structured reporting has been proposed as a solution to improve clinical documentation, while facilitating data reuse for secondary purposes. This study describes the implementation and evaluation of a national standard for tumor board reporting for breast cancer on the basis of the clinical practice guideline and the potential for reusing clinical data for the Netherlands Cancer Registry (NCR).

METHODS Previously, a national information standard for breast cancer was derived from the corresponding Dutch clinical practice guideline. Using data items from the information standard, we developed three different tumor board forms: preoperative, postoperative, and postneoadjuvant-postoperative. The forms were implemented in Amphia Hospital's electronic health record. Quality of clinical documentation and workload before and after implementation were compared.

RESULTS Both draft and final tumor board reports were collected from 27 and 31 patients in baseline and effect measurements, respectively. Completeness of final reports increased from 39.5% to 45.4% (P = .04). The workload for tumor board preparation and discussion did not change significantly. Standardized tumor board reports included 50% (61/122) of the data items carried in the NCR. An automated process was developed to upload information captured in tumor board reports to the NCR database.

CONCLUSION This study shows implementation of a national standard for tumor board reports improves quality of clinical documentation, without increasing clinical workload. Simultaneously, our work enables data reuse for secondary purposes like cancer registration.

Introduction

Tumor boards[1], clinical practice guidelines[2] and cancer registries[3] are intertwined cancer care quality instruments. Tumor boards perform two separate tasks. First, they perform a multidisciplinary review of the patient status during which data previously reported by ancillary services (e.g. Radiology or Pathology) may be aggregated or reinterpreted.[4–7] For example, a tumor board may decide that, for a particular case, a tumor diameter is better approximated on ultrasound than on MRI, which may lead to readjustment of the tumor stage. Subsequently, based on the outcome of the review, the tumor board will recommend a course of action. This final recommendation, together with any -potentially readjusted- findings that drive it, should be documented in the EHR in a Tumor Board Report (TBR).[8]

Clinical practice guidelines are the embodiment of the current status of scientific knowledge and (should) form the basis of the tumor board recommendations.[2] However, the format of most guidelines are far from ideal for consultation at the point of (multidisciplinary) decision making.[9]

Cancer registries, such as the Netherlands Cancer Registry (NCR), collect patient data generated during routine care. They are the basis for epidemiologic outcomes research, results of which are used to evaluate and refine the guidelines used in tumor board decision making.

In the current situation, there are several areas of improvement, being tumor board report quality, clinician workload and data reuse.

Firstly, quality of tumor board reports varies wildly between hospitals, while at the same time it is easily understood that proper documentation may directly influence patient outcomes.[10]

Secondly, although valuable for patient care, tumor board meetings place a burden on physicians' time. Due to the meeting's high pace preparation is required, which consists of summarizing patient history, clinical findings and ancillary information. This is complex work: it requires knowledge and skill to reconstruct a patient's medical timeline from progress notes and to determine what is relevant for the upcoming discussion. After the meeting, it is customary to notify the patient's GP of the outcome by means of a clinical letter. This too places a burden on time.[11]

Thirdly, cancer registries to date either rely on self-reporting by hospitals or employ professionally trained data managers (cancer registrars) to obtain data from medical records.[12] In both cases, this requires significant effort.

Standardized structured reporting (SSR) has been proposed as a solution to support clinicians to produce more complete and consistent documentation.[13, 14] It entails capturing data in discrete fields using (international) terminology systems, like SNOMED. As a result, reuse of information for secondary purposes is facilitated at the same time.[15, 16]

In this study we describe the implementation and evaluation of a national standard for tumor board reporting for breast cancer care, based on the national guideline effected according the method described by Hendriks et al.[17] We investigated the effects of the implementation regarding 1) the quality of clinical documentation with 2) the associated workload and 3) reuse of data for a) the NCR and b) automatic text generation for GP letters. Additionally, we investigated the effect on additional data entry and changes required *during* the tumor board meeting and whether cancer registrars can play a role in supporting tumor board preparation.

Methods

Design

For this study a before-after design was used. Data collection took place before (baseline measurement) and after (effect measurement) implementing the national standard for breast cancer tumor board forms (forms are defined as predefined questionnaires).

The study design was submitted to a medical ethics committee, but was considered exempt from approval.

Definition of the national information standard

Previously, we derived a national information standard for breast cancer from the corresponding Dutch clinical practice guideline (CPG). Briefly, the guideline was analyzed and translated into clinical decision trees.[17] In a decision tree, nodes, branches and leaves represent data-items (patient- or disease characteristics, e.g. "tumor diameter"), values or cut-off points (e.g. "≤ 5 mm") and guideline recommendations (e.g. "perform a lumpectomy") respectively. We encoded the data-items, together with any value sets, using international standards (e.g. SNOMED) where possible. The resulting list of data-items makes up the information standard for breast cancer.

The information standard was approved by the EHR standardization workgroup of the National Breast Cancer Network of the Netherlands (NABON), members of which are surgeons, physicians, radiation oncologists, radiologists, pathologists, clinical geneticists, nuclear medicine specialists and nurses with a formal mandate of their respective Dutch national associations. It is published online (in Dutch).[18]

Implementation

The study was carried out at Amphia Hospital (Breda, the Netherlands), an 837-bed hospital treating 380 new breast cancer cases annually. The multidisciplinary team meeting in the Amphia hospital takes place once a week. There are 5-9 new breast cancer patients presented every week. Here, tumor board preparation is performed by nurse practitioners and consists of collection and entry of relevant patient data in a form in the EHR. During the tumor board, this form is updated and after approval its status changes from "draft" to "final". All related clinical documentation within the EHR is performed by surgical and medical oncologist departments (ancillary departments have their own information systems).

We distinguished three different tumor boards: 1) pre-operative, 2) post-operative, and 3) post-neoadjuvant-post-operative. For each, the standardized and structured tumor board forms were composed using the data-items from the national information standard that are relevant for the different tumor board types (e.g. the field "cT" was included in the pre-operative form where "pT" was included in the post-operative forms) (see Supplement 1).

For the purpose of the study, these forms were implemented in the hospital's electronic health record (EHR, Epic Hyperspace®). EHR functionality was configured for generating full-text clinical notes from entered form data. The generated clinical notes were subsequently reused in correspondence to the patients' GP. End users were trained to work with the standardized forms before introduction into daily clinical practice.

It should be noted that Amphia already used forms in the EHR for structured tumor board reporting at baseline. However, the previous forms were less comprehensive, not associated with any terminology system, not aligned with the national information standard and did not generate full-text clinical notes.

Finally, a data extraction and transformation process was developed to automatically upload the information, captured in the tumor board reports, to the NCR database. Electronic messaging used Health Level 7 – Fast Healthcare Interoperability Resources (HL7 FHIR).

Data collection

Both draft and final tumor board reports, created as part of routine clinical care, were collected in baseline and effect measurements. Cancer registrars prepared tumor board reports in a sandbox environment (a one-day old copy of the production environment) of the EHR.

Additionally, time required for tumor board preparation (by nurse practitioners) and tumor board discussion were measured by manually clocking each case. Prior to the measurements we defined start and stop indications per task. Likewise, it was decided

how to deal with potential interruptions. To check consistency in time measurements the first batch was measured by two researchers.

Finally, a questionnaire, consisting of 25 statements regarding the usability of the tumor board forms, was created. The statements were divided over the domains simplicity, clarity, readability and general impression. Each statement was scored on a four-point scale, with higher values indicating better usability. Questionnaire responses were obtained from the medical professionals (N=4) who were actively involved in tumor board preparations.

Assessing quality of (draft) tumor board reports

The goal of a tumor board report is to reflect the outcome of the multidisciplinary case review and communicate the recommended course of action. As such, it should contain all tumor-board data-items substantiating the final recommendation. For an individual case, this corresponds to a set of data-items (and their values) that make up a path through the decision tree(s) in question.

This also means that the minimally required set of data, varies from case to case. For example, according to the guideline breast cancer, if a patient has metastatic disease, details about the primary tumor or lymph node involvement are irrelevant when selecting primary treatment. However, for non-metastatic disease, these details are required. To complicate matters, in a tree there may be multiple paths leading to a single recommendation. As a result, determining which (and consequently how many) data-items *should* have been reported for individual cases becomes difficult or even impossible in case of missing data (Figure 1).

Quality of a tumor board report was therefore operationalized as follows. Relevant subsets defined previously (see "Definition of the national information standard") were considered the gold standard for each type of tumor board. Completeness was defined as the number of data-items contained in the report, divided by the number of data-items in the relevant subset. Sample size calculations regarding this primary objective indicated a minimal number of 30 reports for demonstrating a statistically significant difference.

Considering it is common that only a few data-items are required to complete a path through a decision tree and thus determine the appropriate guideline recommendation, completeness < 100% is expected and does not indicate a low-quality report (Figure 1). Therefore, the measure should not be used in an absolute sense. However, it can be used to measure changes (e.g. in a baseline and intervention setting).

To additionally measure the impact on quality of tumor board preparation we *a)* compared the scored data-items of the tumor board reports in draft status with their final version and *b)* compared the scored data-items of the drafts prepared by nurse practitioners with those prepared by cancer registrars.

Assessing feasibility of data–reuse for cancer registration

To assess the potential for reusing tumor board data for cancer registries, we determined the overlap between the data-items in the tumor board forms and the data-items currently registered in the NCR for breast cancer.

Statistical evaluation

For all statistical analyses, two-sided unpaired t-tests were used to compare data from before and after implementation. A p-value < 0.05 was considered to be statistically significant. Copy (and adapt if necessary) from paper

Results

Implementation

The national information standard for breast cancer was derived as described in the method section and is composed of 121 data-items, 113 of which were found in the guideline. The number of data-items for 1) pre-operative, 2) post-operative and 3) post-neoadjuvant-post-operative tumor boards forms were 37, 39 and 37, respectively (Table 1).

Standardized structured tumor board forms, automatic text generation for GP letters and HL7 FHIR based message exchange to the NCR were successfully implemented. The FHIR message definitions can be found online on simplifier.net.[19]

Assessing quality of tumor board reports

Clinical documentation

Draft and final tumor board reports were collected from 27 and 31 patients in baseline and effect measurements, respectively (Table 1). Measurements for every patient were performed for 5 subsequent weeks in baseline and in effect setting. Completeness of final tumor board reports increased from 39.5% in baseline to 45.4% in effect measurements ($p = 0.04$) (Table 2).

During the tumor board meeting in baseline on average (14.9–13 =) 1.9 data-items were added to the report. In the effect measurements this delta was (17 – 16.6 =) 0.4 data-items (Table 2). This change was not statistically significant.

At baseline, when comparing values of individual data-items between the draft and final report, 26 out of 414 (6.3%) were documented with different values. In the effect measurements, different values were recorded in 22 out of 531 (4.1%) data-items (Table 3). This change was not statistically significant.

Tumor board preparation by cancer registrars

Cancer registrars prepared 16 and 23 cases in baseline and effect measurements respectively (Table 1). When comparing draft tumor board reports prepared by nurse practitioners and cancer registrars, no statistically significant difference was found regarding completeness in baseline and effect measurements (Table 2).

In 16 (draft) tumor board reports in the baseline measurement, an absolute number of 210 data-items were scored either by nurse practitioners or cancer registrars. Out of these, 3 (1.4%) data-items were recorded by nurse practitioners only, and 11 (5.2%) data-items by cancer registrars only. 173 (82.4%) data-items were recorded by both with equal values, and 23 (11.0%) were recorded with discordant values.

Similarly, in the effect measurement 381 data-items were scored across 23 (draft) tumor board reports. Here, 14 (3.9%) data-items were only scored by nurse practitioners, 46 (12.7%) were only scored by cancer registrars, 286 (79.2%) data-items were documented by both with corresponding values and 15 (4.2%) data-items were recorded with disagreeing values (Table 4).

Workload

Mean time involved with tumor board preparation by nurse practitioners did not change significantly (from 4:06 minutes [SD=1:44] to 4:39 minutes [SD=1:59], $p = 0.28$). Time for tumor board discussion per patient did not change significantly (from 2:19 minutes [SD=1:27] to 2:43 minutes [SD=1:41], $p = 0.35$) (Table 5).

Assessing feasibility of data reuse for cancer registries

Standardized tumor board reports included 50% (61/122) of the data-items carried in the NCR (Supplement 1).

End user satisfaction

The mean overall usability score (range 1-4) was 2.63 in baseline and 2.84 in effect measurement, suggesting an overall improvement. Distributed over the subdomains the mean scores were: simplicity: 2.75 and 2.75, clarity: 2.50 and 2.96, readability: 2.71 and 3.00, and general impression 2.69 and 3.22 in baseline and effect measurements, respectively.

Discussion

This study shows that implementation of a national standard for tumor board reports, improves quality of clinical documentation and is possible without increasing the clinical workload. At the same time, our work enables data reuse for secondary purposes like cancer registration.

Potential downsides of structured reporting may include perceived disproportionate burden to physician workload and limitations in reporting freedom.[20] Based on the results on user satisfaction our study did not corroborate these presumed downsides.

However, as mentioned in the Methods section, it should be noted that Amphia already used structured EHR forms for tumor board reporting at baseline. Compared to coming from a completely free-text baseline, this could have led to underestimation of the extra effort required for structured reporting, but might also have led to underestimation of the improvement in clinical documentation quality. Our results are comparable to those in pathology where a national standard for structured reporting also improved completeness of clinically relevant data.[16]

The completion rates of approximately 50% may seems low but need to be interpreted carefully. Indeed, it can be partly explained by certain data-items not being filled in in clinical practice. Yet as demonstrated by the logic tree in Figure 1, on a per case basis this low completion does not imply a similar amount of data-items required for guideline based treatment decisions were missing.

Another reason for lower completion may reside in usability issues related to EHR systems. Indeed there are recent studies showing a relation between physician burn out and registration burden and EHR usability.[21] The degree to which usability can be taken into account when implementing an information standard is limited by the possibilities of the EHR system. Measures that were taken to minimize frustration were that no items were considered mandatory, taken into account that in clinical practice cases do occur where information simply is not available. Secondly, the forms to a degree allow to hide information which was not required in specific cases. For example, detailed information required for a lesion, are only shown to the user if a lesion is actually present.

With respect to data reuse for cancer registries, approximately 50% of data-items defined in tumor board forms are currently carried in the NCR. The actual degree of reuse potential depends on the completeness of the tumor board reports. Despite the improvement in documentation quality, we observed an overall low completeness of tumor board reports. This may limit ability for data reuse in practice. Low completeness could be partially explained by not documenting negative findings (e.g. not explicitly documenting "M0" in absence of metastatic disease). Lack of disciplined use of structured reporting forms by clinicians is a well-known phenomenon often attributed to poor EHR usability and the aforementioned (perceived) disproportionate burden and limitations in reporting freedom.[22]

We investigated the possibility to have cancer registrars assist in tumor board preparation. As there were no significant differences in completeness and only a limited number of discrepancies in scored data-items between draft tumor board forms prepared

by nurse practitioners and cancer registrars, the results hint towards the possibility of employing cancer registrars for this purpose. In absence of a ground truth, we were not able to evaluate discrepancies in documentation between nurse practitioners and cancer registrars and establish which party was correct (if any). However, the number of data-items scored with different values decreased from 11% in baseline to 4.1% in effect measurements.

This change was largely explained by the fact that the (introduced) standardized tumor board forms prevented scoring values that were not part of the official TNM-classification system (e.g. cMX). To evaluate the remaining discordant values and get a better understanding of the cancer registrars drafts, further investigation would be recommended.

To evaluate another avenue that might reduce tumor board preparation workload, we estimated the amount of information in the tumor board form that is generated by ancillary services, like radiology and pathology (Supplement 1). This suggested that 59% of the data-items in the forms could be automatically pre-populated if supported by the underlying technology. Actual benefits, like with reuse of data from the tumor board report, depend on the degree to which ancillary departments report required information.

Evaluation regarding usability in the effect measurement showed an improvement over the baseline, although the number of participants in the survey was low. This was partially due to the fact that only a limited number of clinicians is actively involved in the information management tasks surrounding tumor board meetings. The survey results were consistent with the positive personal feedback we received from these clinicians.

EHR implementations of structured reporting are not unique, but usually based on local physician preferences. The strength of our approach is that it's a) based on a national standard, that b) is derived from the national guideline and c) enables evaluation of guideline adherence. As such, this study provides a roadmap for tumor board meetings for other tumors than breast cancer.

Tables & Figures

	Number of items in tumor board form	Number of draft & final reports collected		Number of draft reports prepared by data managers	
		Baseline	Effect	Baseline	Effect
Preoperative	37	14	17	5	14
Postoperative	39	11	8	10	5
Post-neoadjuvant-postoperative	37	2	6	1	4
Total	113	27	31	16	23

Table 1: Overview of the number of data-items per tumor board form and number of tumor board reports collected. The draft reports that were subsequently finalized, were prepared by nurse practitioners. Draft reports prepared by data managers were not involved in the clinical process.

	Draft				Final	
Author	Nurse practitioner		Data manager		Tumor board	
Type	Baseline	Effect	Baseline	Effect	Baseline	Effect
Preoperative	16 (43%)	18.5 (50%)	18.4 (50%)	18.2 (49%)	18.2 (49%)	18 (50%)
Postoperative	9.8 (25%)	13.5 (35%)	14.3 (37%)	15.2 (39%)	15.2 (39%)	15 (38%)
Post-neo-adjuvant-postoperative	10 (27%)	15.2 (41%)	12 (32%)	16 (43%)	16 (43%)	15.5 (42%)
Average	13.0 (34.7%)	16.6 (44.3%)	15.8 (43.1%)	17.2 (45.3%)	14.9 (39.5%)	17.0 (45.4%)

Table 2: Mean number of data-items scored with completeness (in parentheses), in the draft and final tumor board reports, for each type of tumor board report.

	Number of items added by tumor board	Number of items changed by tumor board	Number of items unchanged	Number of items total
Baseline				
Preoperative	14	12	195	221
Postoperative	65	14	90	169
Post-neoadjuvant-postoperative	12	0	12	24
Total	91 (22.0%)	26 (6.3%)	297 (71.7%)	414 (100%)
Effect				
Preoperative	14	11	300	325
Postoperative	13	5	102	120
Post-neoadjuvant-postoperative	8	6	72	86
Total	35 (6.6%)	22 (4.1%)	474 (89.3%)	531 (100%)

Table 3: Overview of changes between draft (prepared by nurse practitioners) and final tumor board reports, for each type in baseline and effect measurements. The first column shows the number of data-items added to the draft by the tumor board. The second and third column display the number of items that were changed and remained unchanged, respectively.

	Scored by data managers only	Scored by nurse practitioners only	Scored with different values	Scored with corres-ponding values	Total
Baseline					
Preoperative	4	1	8	61	74
Postoperative	6	0	15	104	125
Post-neoadjuvant-postoperative	1	2	0	8	11
Total	11 (5.2%)	3 (1.4%)	23 (11.0%)	173 (82.4%)	210 (100%)
Effect					
Preoperative	26	8	11	190	241
Postoperative	10	0	0	56	66
Post-neoadjuvant-postoperative	10	6	4	40	61
Total	46 (12.7%)	14 (3.9%)	15 (4.2%)	286 (79.2%)	361 (100%)

Table 4: Comparison of draft tumor board reports as prepared by data managers and nurse practitioners, for each type of tumor board in baseline and effect measurements. The first two columns show the number of items that were either documented by data managers or by nurse practitioners. The third and fourth columns show the number of items they disagreed and agreed on, respectively.

Activity	Baseline [SD]	Effect [SD]	Delta
Tumor board preparation	00:04:06 [1:44]	00:04:39 [1:59]	+00:00:33
Tumor board discussion	00:02:19 [1:27]	00:02:43 [1:41]	+00:00:24

Table 5: Mean time (in minutes) spent per patient for tumor board related activities in baseline and effect measurements. SD = standard deviation.

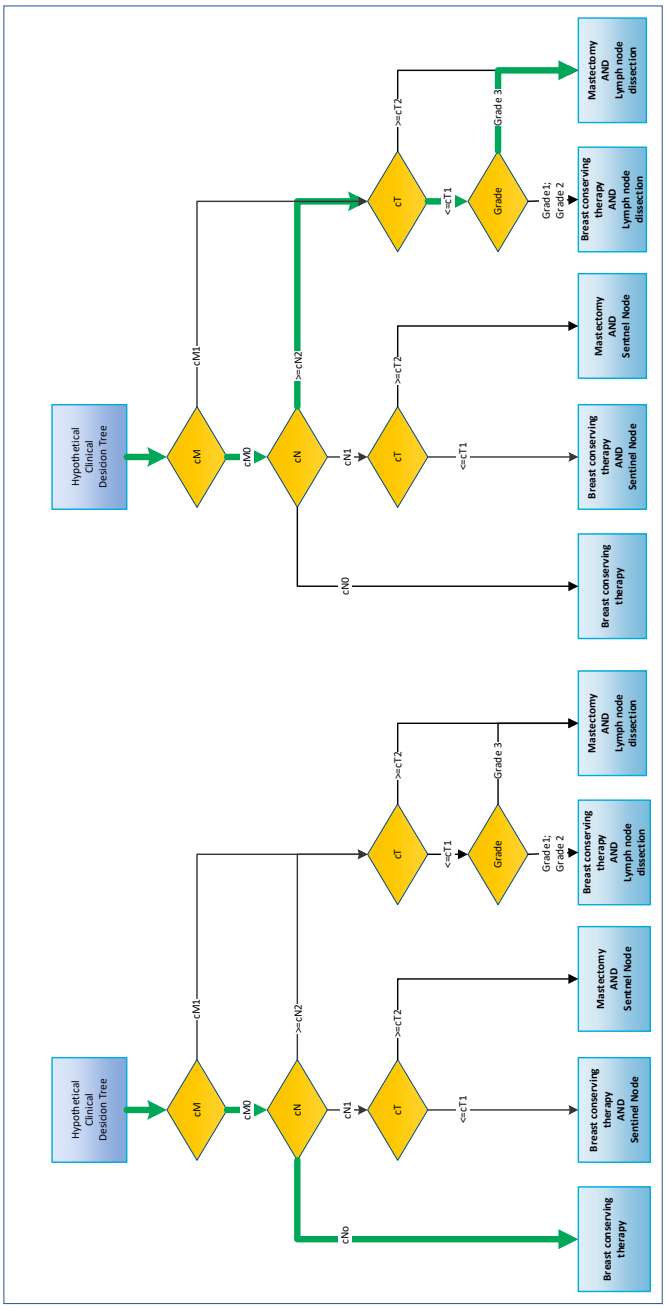


Figure 1: Example of a hypothetical Clinical Decision Tree (CDT). The nodes (yellow) represent patient- and disease characteristics, the branches represent values of these characteristics and the leaves at the bottom (blue) contain recommendations. Every patient runs through the CDT (top-down) on a single, individual path, passing a selection of the characteristics leading to a recommendation. As indicated by the green paths, on the left panel two data-items are required to be provided with a recommendation, on the right panel this is four.

References

1. Lamb BW, Green JSA, Benn J, et al: Improving decision making in multidisciplinary tumor boards: Prospective longitudinal evaluation of a multicomponent intervention for 1,421 patients. J Am Coll Surg , 2013
2. Institute of Medicine: Clinical Practice Guidelines We Can Trust. 2011
3. Parkin DM: The role of cancer registries in cancer control. Int J Clin Oncol 13:102–111, 2008
4. Newman EA, Guest AB, Helvie MA, et al: Changes in surgical management resulting from case review at a breast cancer multidisciplinary tumor board. Cancer 107:2346–2351, 2006
5. Thenappan A, Halaweish I, Mody RJ, et al: Review at a multidisciplinary tumor board impacts critical management decisions of pediatric patients with cancer. Pediatr Blood Cancer , 2017
6. Van Hagen P, Spaander MCW, Van Der Gaast A, et al: Impact of a multidisciplinary tumour board meeting for upper-GI malignancies on clinical decision making: A prospective cohort study. Int J Clin Oncol , 2013
7. Wheless SA, McKinney KA, Zanation AM: A prospective study of the clinical impact of a multidisciplinary head and neck tumor board. Otolaryngol - Head Neck Surg , 2010
8. Pillay B, Wootten AC, Crowe H, et al: The impact of multidisciplinary team meetings on patient assessment, management and outcomes in oncology settings: A systematic review of the literature. Cancer Treat Rev , 2016
9. Peleg M, Tu S, Bury J, et al: Comparing computer-interpretable guideline models: a case-study approach. J Am Med Inform Assoc 10:52–68, 2003
10. Darko-Yawson S, Ellingsen G: Assessing and Improving EHRs Data Quality through a Socio-technical Approach, in Procedia Computer Science. 2016
11. El Saghir NS, Keating NL, Carlson RW, et al: Tumor boards: optimizing the structure and improving efficiency of multidisciplinary management of patients with cancer worldwide. Am Soc Clin Oncol Educ book Am Soc Clin Oncol Annu Meet e461-6, 2014
12. Schouten LJ, Jager JJ, van den Brandt PA: Quality of cancer registry data: a comparison of data provided by clinicians with those of registration personnel. Br J Cancer 68:974–977, 1993
13. Farrugia DJ, Fischer TD, Delitto D, et al: Improved Breast Cancer Care Quality Metrics After Implementation of a Standardized Tumor Board Documentation Template. J Oncol Pract , 2015
14. Woods YL, Mukhtar S, McClements P, et al: A survey of reporting of colorectal cancer in Scotland: Compliance with guidelines and effect of proforma reporting. J Clin Pathol , 2014
15. Renshaw AA, Mena-Allauca M, Gould EW, et al: Synoptic Reporting: Evidence-Based Review and Future Directions. JCO Clin Cancer Informatics , 2018
16. Sluijter CE, van Lonkhuijzen LRCW, van Slooten H-J, et al: The effects of implementing synoptic pathology reporting in cancer diagnosis: a systematic review. Virchows Arch 468:639–649, 2016
17. Hendriks MP, Verbeek XAAM, van Vegchel T, et al: Transformation of the National Breast Cancer Guideline Into Data-Driven Clinical Decision Trees. JCO Clin Cancer Informatics , 2019
18. NABON: Information Standard Breast Cancer [Internet]Available from: https://decor.nictiz.nl/art-decor/decor-project--onco-mamma-
19. NABON: Breast Cancer Tumor Board Questionnaires [Internet]. Simplifier.net, 2017 Available from: https://simplifier.net/breastcancertumorboardquestionnaires/~resources.
20. Swillens JEM, Sluijter CE, Overbeek LIH, et al: Identification of barriers and facilitators in nationwide implementation of standardized structured reporting in pathology: a mixed method study. Virchows Arch , 2019
21. Brown T, Bergman S: Doctors, Nurses and the paperwork crises that could unite them [Internet]. New York Times , 2019Available from: https://www.nytimes.com/2019/12/31/opinion/doctors-nurses-and-the-paperwork-crisis-that-could-unite-them.html
22. Gawande A: Why Doctors Hate Their Computers [Internet]. New Yorker - Ann Med , 2018Available from: https://www.newyorker.com/magazine/2018/11/12/why-doctors-hate-their-computers

Supplementary Materials

Supplement 1. The implemented TBR forms, their included data-items and their presence in the NCR and primary source reports.

Data-items (population/disease characteristics and workflow data)	Pre-operative TBR	Post-operative TBR	Post-neoadjuvant-postoperative TBR	NCR*	Primary source reports
Tumor board date*	X	X	X	X	Administrative
Tumor board report approval date*	X	X	X		Administrative
Present specialists*	X	X	X		Administrative
Clinical question*	X	X	X		Administrative
Pregnancy	X	X			Clinical
Disease laterality	X	X	X	X	Clinical
Disease localization	X	X	X	X	Clinical
BI-RADS	X			X	Radiology
Calcifications in breast	X				Radiology
Solid mass in breast	X				Radiology
Tumor laterality	X			X	Radiology
Tumor localization	X			X	Radiology
cT-diameter	X			X	Radiology
cT4 characteristics	X				Radiology
Morphology	X	X	X	X	Pathology
Grade	X	X	X	X	Pathology
cT	X			X	TBR
Tumor distribution	X			X	Radiology
Solid component on mammography	X				Radiology
ER-percentage	X	X	X	X	Pathology
HER2 status	X	X	X	X	Pathology
Lymph node laterality	X	X		X	Radiology
Fixed axillary lymph nodes	X				Radiology
Clinical positive axillary lymph nodes	X				Radiology
Clinical positive parasternal lymph nodes	X	X	X		Radiology
Clinical positive infraclavicular lymph nodes	X				Radiology
Clinical positive supraclavicular lymph nodes	X				Radiology
cN	X			X	TBR
Localization metastasis	X			X	Radiology
ER-percentage metastases	X				Pathology

Data-items (population/disease characteristics and workflow data)	Pre-operative TBR	Post-operative TBR	Post-neoadjuvant-postoperative TBR	NCR#	Primary source reports
HER2 status metastases	X				Pathology
cM	X			X	TBR
cTNM	X			X	TBR
Clinical study advice*	X	X	X		TBR
Clinical study*	X	X	X	X	TBR
Policy*	X	X	X	X	TBR
Motivation for CPG deviation*	X	X	X		TBR
pT diameter		X		X	Pathology
pT4 characteristics		X			Pathology
Cutting edge		X	X	X	Pathology
Angio invasion		X	X	X	Pathology
pT		X		X	Pathology
Positive supraclavicular lymph nodes in SN		X	X		Pathology
Positive parasternal lymph nodes in SN		X	X		Pathology
Positive infraclavicular lymph nodes		X	X		Pathology
Number of axillary lymph nodes with micro or macro metastases in SN		X	X	X	Pathology
Number of axillary lymph nodes with macro metastases in SN		X	X	X	Pathology
Number of axillary lymph nodes with micro metastases in SN		X	X	X	Pathology
Number of axillary lymph nodes with isolated tumor cells in SN		X	X	X	Pathology
Extra nodal growth in SN		X	X	X	Pathology
Extra nodal growth in lymph node dissection		X	X	X	Pathology
Positive axillary top lymph node		X	X	X	Pathology
pN		X		X	Pathology
pM		X		X	Pathology
pTNM		X		X	TBR
Genetic mutation		X	X	X	Pathology
N0 risk status		X			Pathology
First re-excision		X	X	X	TBR
Lymph node surgery		X	X	X	TBR
ypT 4 diameter			X	X	Pathology

Data-items (population/disease characteristics and workflow data)	Pre-operative TBR	Post-operative TBR	Post-neoadjuvant-postoperative TBR	NCR#	Primary source reports
ypT 4 characteristics			X		Pathology
Pathological response primary tumor			X	X	Pathology
ypT			X	X	Pathology
ypN			X	X	Pathology
ypM			X	X	Pathology
ypTNM			X	X	TBR
Total	37	39	37	61#	

* = additional data items (not CPG) identified by care professionals in order to optimally support the tumor board processes; # = several NCR data-items are registered multiple times throughout the clinical pathway; NCR = Netherlands Cancer Registry; TBR = Tumor Board Report

3

Predicting lung cancer survival using probabilistic reclassification of TNM-editions with a Bayesian network

Adapted from Sieswerda M.S., Bermejo I., Geleijnse G., Aarts M.J., Lemmens V.E.P.P., De Ruysscher D., Dekker A.L.A.J., Verbeek X.A.A.M.; "**Predicting Lung Cancer Survival Using Probabilistic Reclassification of TNM Editions With a Bayesian Network**"; *JCO Clin Cancer Inform*. 2020;4:436-443. doi:10.1200/CCI.19.00136

Abstract

PURPOSE: The TNM classification system is used for prognosis, treatment, and research. Regular updates potentially break backward compatibility. Reclassification is not always possible, is labor intensive, or requires additional data. We developed a Bayesian network (BN) for reclassifying the 5th, 6th, and 7th editions of the TNM and predicting survival for non-small-cell lung cancer (NSCLC) without training data with known classifications in multiple editions.

METHODS: Data were obtained from the Netherlands Cancer Registry (n = 146,084). A BN was designed with nodes for TNM edition and survival, and a group of nodes was designed for all TNM editions, with a group for edition 7 only. Before learning conditional probabilities, priors for relations between the groups were manually specified after analysis of changes between editions. For performance evaluation only, part of the 7th edition test data were manually reclassified. Performance was evaluated using sensitivity, specificity, and accuracy. Two-year survival was evaluated with the receiver operating characteristic area under the curve (AUC), and model calibration was visualized.

RESULTS: Manual reclassification of 7th to 6th edition stage group as ground truth for testing was impossible in 5.6% of the patients. Predicting 6th edition stage grouping using 7th edition data and vice versa resulted in average accuracies, sensitivities, and specificities between 0.85 and 0.99. The AUC for 2-year survival was 0.81.

CONCLUSION: We have successfully created a BN for reclassifying TNM stage grouping across TNM editions and predicting survival in NSCLC without knowing the true TNM classification in various editions in the training set. We suggest binary prediction of survival is less relevant than predicted probability and model calibration. For research, probabilities can be used for weighted reclassification.

Introduction

In cancer care, the TNM system for classification of malignant tumors guides treatment decisions, aids in stratifying patients for research and helps clinicians assess prognosis.[1,2] This is done by classifying characteristics of the tumor (T-descriptor), local lymph nodes (N-descriptor) and distant metastases (M-descriptor). These descriptors can subsequently be used to compute a stage grouping, essentially summarizing the information.

The system is revised on a regular basis (every 5.7 years on average), during which changes are made both to the individual descriptors and to the stage grouping.[3] Revisions incorporate new developments that improve (outcome) stratification and prognostic capabilities, keeping the classification system relevant. However, because categories can be added or removed, classes with the same label are not necessarily equivalent across editions.[4] Recommendations for care interventions from literature and clinical trials may be based on specific editions of the classification system and it is not always immediately clear how to apply these recommendations to patients classified with a different edition.[5] Also, scientific analysis of patient cohorts classified with different editions must consider the differences across editions.

This issue can be tackled by either mapping class labels from source to target edition or by (re)classifying the patient using the target edition. If mapping is not feasible, additional data is required to help determine the individual descriptors in the target edition. In practice, this process is complicated and these data are usually excluded from the analyses or an approximate mapping is assumed.

For both situations above, it would be helpful to have a model that can aid in the reclassification across TNM-editions. In this paper we develop such a model based on Bayesian Networks (BNs).

BNs are a type of probabilistic graphical model that use a directed acyclic graph (DAG). The nodes represent variables and the directed edges signify (preferably causal) relationships.[6,7] Each node is associated with a probability distribution that is conditional on its parents (i.e. the set of nodes that have a directed edge to that node), which can be written as $P(X \mid Pa_X)$. This leads to a set of (conditional) probability distributions that together define a joint probability distribution, which can be written as $P(X_1, \ldots, X_n) = \prod_{i=0}^{n} P(X_i \mid Pa_{X_i})$. For nodes that are associated with a discrete probability distribution the distribution is defined as a conditional probability table (CPT).

BNs can be used to estimate the probability distribution of a variable given evidence. In contrast to other models, such as logistic regression, BNs do not have dedicated inputs or outputs. Instead, setting evidence on any node updates probabilities throughout the network.

The resulting models can easily be evaluated by medical specialists: the graphical nature of BNs makes interpretation of relationships straightforward and conditional probability aligns well with physicians' reasoning. This is a benefit over black-box approaches such as artificial neural networks or deep learning.

In this work we hypothesized BNs can be used to reclassify data across TNM-editions for non-small cell lung cancer (NSCLC) and predict survival. We further hypothesized such a BN can be learned without knowing the true TNM classification in various editions in the training set by leveraging the correlation between TNM and survival.

Methods

Data

Data were obtained from the population-based Netherlands Cancer Registry (NCR) after approval by the NCR Privacy Review Board and did not require approval from an ethics committee in the Netherlands. The NCR is maintained since 1989 and populated by trained data managers. It contains all cancer occurrences in the Netherlands. Coding is based on international rules and standards. The edition of the TNM classification used depends on the incidence year (Table 1).

Inclusion criteria were, pathology confirmed, ICD-O-3 topology code C34 (bronchus and lung), morphology codes appropriate for NSCLC and incidence year between 1999 and 2016; 2017-2018 were excluded due to insufficient follow-up. Patients with multiple primary tumors were kept since lung cancer is generally dominant in determining survival. A total of 146.084 patients fulfilled our criteria. We obtained the following variables: clinically staged T-, N- and M-descriptors, incidence year, days of follow-up, and vital status at follow-up. A new variable was added to each record specifying the TNM-edition used, based on year of incidence. Survival time, counted from initial diagnosis, was discretized into 5 categories frequently used in literature and clinical practice (Table 2). Discretization introduced NAs for patients that had a follow-up of less than 2 years and were alive at the time of follow-up. The dataset was randomly split into a training set (80%, n=116.858) and a test set (20%, n=29.226). Since the distribution of the different TNM-editions was unequal, the training set was resampled (with replacement) to contain 45.000 samples for each edition. NAs were removed from the test set.

Network structure definition

A BN was designed to predict the 5th and 6th TNM-edition with variables from TNM7 and vice versa, visualized in Figure 1 using BayesiaLab 8.[8] This required *two* sets of four nodes corresponding to T-, N- and M-descriptors, and TNM-stage grouping: one set for input and one for output. By adding two nodes for TNM-edition and survival, a total of 10 nodes was obtained. Relationships between nodes were established to indicate

causal effect. Nodes edition, T_567, N_567, M_567, TNM_567 and death were associated to variables in the dataset, leaving T_7, N_7, M_7 and TNM_7 as hidden nodes. CPTs for the hidden nodes were estimated on a subset of the training set containing only 7th edition data. Relationships between hidden nodes and their observed counterparts were primed by manually estimating CPTs $P(T\_567 \mid edition, T\_7)$, $P(N\_567 \mid edition, N\_7)$ and $P(M\_567 \mid edition, M\_7)$ through analysis of the differences between editions 6 and 7 in the AJCC staging manual (Supplemental Tables Table S5, Table S6, and Table S7); there were no changes between editions 5 and 6, survival differed only marginally (Supplemental Figure S1). Finally, Expectation-Maximization-learning was used to estimate the CPTs of the full network.[9]

Reclassifying TNM7 data as ground truth for testing

Evaluating network reclassification performance requires a ground truth. Therefore, a subset of the test set was manually reclassified as follows.

The additional parameters multifocal, tumor size and sulcus superior were obtained from the NCR for patients in the dataset (Table 2). Since collection of tumor size and sulcus superior involvement started in 2015, the subset was limited to include patients diagnosed in 2015 and 2016 (n=3544).

By comparing definitions for the T, N and M descriptors in the AJCC staging manuals, reclassification rules were defined (Supplemental Table S8). If a one-to-one mapping was impossible, determining a range of values was attempted. With the individually reclassified descriptors, the TNM-stage group was computed.

Evaluating predictive network performance

The network's performance was evaluated on 1) predicting the 6th edition stage grouping 2) predicting the 7th edition stage grouping both using the ground truth set and 3) predicting survival using the test set. Predicting the 5th edition was not evaluated separately, since the 5th and 6th edition are equivalent.

The HUGIN Analysis Wizard was used to compute the confusion matrix for each evaluation by selecting the state with the highest belief as predicted state.

Macro- and micro-averaging are used in multi-class problems to combine multiple metrics into single values (see Supplemental "Calculating micro- and macro-averages").[10] Micro- and macro-averaged sensitivity, specificity and accuracy were computed with the Python programming language.[11–14] We additionally used BayesiaLab to calculate the receiver operating statistic area under the curve (ROC-AUC) for 2-year survival.

Model calibration for survival was visualized using a bubble-plot and curve. Briefly, each unique combination of inputs (i.e. edition, T_567, N_567, M_567 and TNM_567)

defines a subpopulation. The plots show, for each subpopulation, the relation between a survival category's predicted probability and its observed frequency in the dataset. The size of each bubble indicates the population's size in the dataset. The calibration curve, computed using scikit-learn in Python, averages predicted and observed values by applying quantile-based binning with 1000 samples per bin (21 bins).[14,15]

Results

Network structure & parameters

The final network structure is shown in Figure 1. Each node represents a random variable and shows its state's prior probabilities in a histogram. The arrows indicate (causal) relations between the nodes. A node's underlying CPT is conditional on its parents.

We assumed that the newest available TNM-edition approximates the true disease state best and thus has a causal relation with survival. This also means, since the TNM-edition used for staging a patient should not (causally) influence survival, each of the nodes T_567, N_567 and M_567 has two parents: edition and its corresponding unobserved/hidden counterpart.

After the network structure was created, EM-learning was successfully applied. Percentages shown on each node show the prior probabilities for each state. In general, the model behaved as expected: an increase in tumor stage in any of the T-, N-, M- and TNM-nodes corresponded to poorer survival. Additionally, the relations between the hidden and the observed TNM-nodes were close to the priors set manually.

Manual reclassification of the dataset

Multifocality, tumor size and sulcus superior involvement were not always available for all patients. For 150 records (4%), the T-descriptor could not be determined due to these missing values. 451 cases (13%) were multifocal T4 tumors in the 7th edition for which it was not possible to determine the corresponding 6th edition T-descriptor without additional information. Another 216 records (6%) could only be classified into a range (e.g. [T2 - T3]). As a result, for a total of 817 records (23%) the T-descriptor could not be completely reclassified (Supplemental Figure S2). The N-descriptor could be reclassified in all cases. There were 85 multifocal T4 tumors (2%) that did not have distant metastases in the 7th edition, making it impossible to determine the M-descriptor in TNM6.

Computation of the stage group does not always require the T- or N-descriptor. For example, if a patient has distant metastases (M1), the stage group will always be 4. Also, in patients without distant metastases (M0), if the lymph node metastases are classified as N2, any T2 or T3 will yield a stage group of 3A.

Consequently, in 3344 out of 3544 records (94.4%) it was possible to definitively determine the stage group. In the remaining 200 cases (5.6%), only a range or a set of stage groups could be identified. 52 cases were either stage 1B or 2B, 13 stage 2B or 3A, 85 stage 3A or 4. In 20 cases the stage could range between 1A and 3B. Finally, in 6 cases the stage could range between 2A and 3B (Supplemental Figure S3). Partial classifications were removed from the ground truth set.

Evaluating predictive network performance

Predicting 6th edition stage grouping using 7th edition data resulted in an average accuracy of 0.99. Macro- and micro-sensitivity were 0.92 and 0.96 respectively. Macro- and micro-specificity were 0.99 and 0.99 respectively. Inspection of the confusion matrix showed two situations where misclassification was notable. In 41% of the cases that were stage 2B were misclassified as stage 1B. In 21% of the cases stage 3B was misclassified as stage 1B, 2B or 3A. See also Supplemental "In-depth analysis of frequent misclassifications".

Predicting 7th edition stage grouping using 6th edition data yielded an average accuracy of 0.99. Macro- and micro-sensitivity were 0.85 and 0.95, while macro- and micro-specificity were 0.99 and 0.99 respectively. Again, inspection of the confusion matrix showed two notable situations where misclassification was apparent. Stage 2A was misclassified as 1A in 53% and stage 2B was misclassified as 2A or 3A in 59% of the cases. See also Supplemental "In-depth analysis of frequent misclassifications".

Macro- and micro- sensitivity, -specificity and accuracy for survival are listed in Table 3.

The confusion matrices underlying these statistics can be found as a supplement (Supplemental Tables Table S9, Table S10 and Table S11 respectively) together with the sensitivity, specificity and ROC-AUC for predicting each stage-group (Supplemental Table S12) and the corresponding set of ROC-AUC curves (Supplemental Figure S4 and Figure S5).

The ROC-AUC for ≥ 2-year survival was 0.81, determined using BayesiaLab and is shown in Figure 2. Model calibration was computed and is shown in Figure 3.

Discussion

Mapping class labels requires that each label in the source edition (classification system) can be translated to a label in the target edition. Generally, this is not possible going from coarser to more granular classification systems (e.g. from the 6th to the 7th edition of the TNM classification system). In these situations, additional variables are required: either for use *in conjunction with* the original classification or to fully stage a tumor in the target edition. These additional variables are frequently unavailable or obtaining them comes with great cost.

We experienced these issues firsthand. When creating the ground truth for evaluating predictive performance, only a relatively small subset of the total test set (12%, patients with a year of diagnosis between 2015-2016) could be considered for manual, rule-based reclassification. Even then, reclassification of the T-descriptor was not possible in 23% of the cases in the subset. Specifically, we could not determine when a T3 in TNM7 would be a T2 in TNM6 as this required information about the presence of invasion into nearby anatomical structures. Similarly, we couldn't determine the 6th edition T- and M-descriptors for 7th edition non-metastatic, multifocal T4 tumors since this required knowledge of the location(s) of the additional tumor nodules. Therefore, we did not evaluate performance of predicting individual descriptors.

On a more aggregated level, we couldn't fully determine the 6th edition stage group in 200 cases (5.6%). Although a relatively small number or records was involved, we thought this might still bias the test set. To investigate the potential effect of this bias, we performed an additional analysis. We looked at the (in real-life impossible) worst-case scenario by assuming that all 174 records where we could not decide between two stages (e.g. stage 1B or 2B) each record would *always* be incorrectly classified, essentially doubling the error. After modifying the confusion matrices in this way, the averaged statistics were recalculated. Micro-/macro-averaged specificity and accuracy were almost unaffected (values changed from 0.99 and 0.99 to 0.98 and 0.96 respectively, i.e. a maximum decrease of 0.03). Micro-/macro-averaged sensitivity decreased by a maximum of 0.1.

When using the ground truth set, the BN performed very well when reclassifying the TNM-stage group between editions: when predicting the 6th edition to the 7th edition data all aggregated statistics (sensitivity, specificity and accuracy) were ≥ 0.91. Also, the fact that macro- and micro-averaged statistics were close together, implies the model is relatively insensitive to class imbalances. A similar observation can be made for predicting the 7th edition using 6th edition data. The results were only slightly worse, which is to be expected considering the network has to predict a more granular output from a coarser input. Still, the model had an accuracy of 98%.

Performance for predicting 2-year survival can be considered more than adequate, especially considering the limited number of variables used for making the prediction

and the number of possible outcome-classes. Adding variables could help, but would come with additional complexity of the model. However, clinical decisions are not just based on the most likely outcome: probability is considered as well. Therefore, binary prediction (e.g. 2-year survival: yes/no) seems inadequate to support decision making. Moreover, using a BN in such a way, ignores one of a its major strengths: the fact that it can communicate (conditional) probabilities. The calibration curves show that the model is well calibrated and that the estimated probabilities are close to the real probabilities.

Inspection of the confusion matrix reveals four situations where the model had difficulty in predicting the correct stage group, all related to possible stage-shifting. When predicting the 6th edition stage group from 7th edition data, the errors stem from difficulty in handling cases where the input (node T_7) is T3. Upon inspection of the BN, it seems the model predicts a T3 in the 7th edition to be a T2 with 58% and a T3 with 35% probability in the 6th edition. This explains the majority of mistakes made when the predicted descriptor should have been T3. The original priors for this relation were set close to 50-50 (i.e. a T3 becoming a T2 or T3 with equal probability) before applying EM-learning, so the change in probabilities appears to be an effect of optimizing the relationship between nodes T_567 and death. Additionally, the BN estimates the probability of a T3 in TNM7 being multifocal and thus a T4 in TNM6 to be fairly small at 4.2%. Even if the actual percentage in the dataset is larger, the value T4 (in TNM6) would never be predicted, since the most probable outcome was selected as prediction.

In predicting the 7th edition stage group from 6th edition data, most of the misclassifications can be explained by the observation that a TNM6 T2 becomes either a T2a, T2b or T3. Without additional information (i.e. tumor diameter), it is not possible to be 100% accurate. Similarly, a T4 in TNM6 can become a T3 or T4, depending on multifocality of the tumor.

Even when using training data with classifications in both 6th and 7th editions, the BN would not have been able to make any of these distinctions with certainty, since additional information is needed. However, like with predicting survival, the probabilistic reclassification we applied does not need to yield a discrete result: it is possible to assign a probability to each possible outcome, essentially creating a weighted reclassification. This is especially useful when reclassifying large datasets like the NCR.

We conclude that we have successfully created a BN that can aid in determining the TNM stage group of the 6th edition using 7th edition data and vice versa, by using a training set that does not hold the known classifications in multiple editions but does hold survival to aid in the classification. Knowledge about changes between the editions of the classification system was successfully incorporated by modelling these changes as priors in the CPTs. The model parameters were estimated from data and therefore depend on specific distributions found in the Netherlands. However, considering NSCLC

diagnostics, treatment and survival are comparable in western/developed countries, we expect the BN can be applied here as-is, although validation would be required. This process is likely to work for other tumors and/or editions of the TNM-classification system, but additional research is needed to establish generalizability.

Tables & Figures

Figures

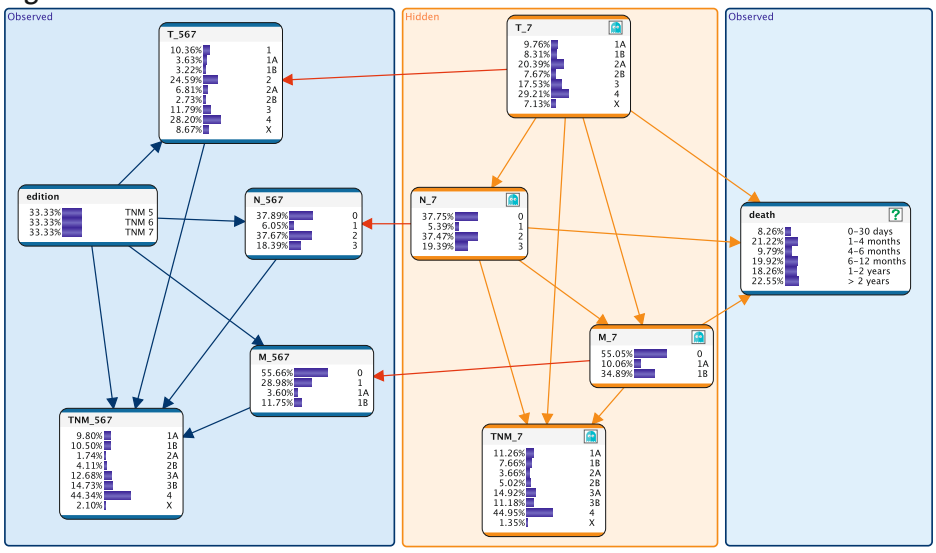


Figure 1: BN (after EM-learning) visualized using BayesiaLab 8. Each node represents a random variable with a (conditional) probability table and shows its states' prior probabilities in a (rotated) histogram. The arrows indicate the (causal) relations between the nodes.The "_567"-suffixed nodes indicate nodes that, in conjunction with the "edition" node, can take on values from all TNM-editions. The "_7"-suffixed nodes can take on 7th edition values only.

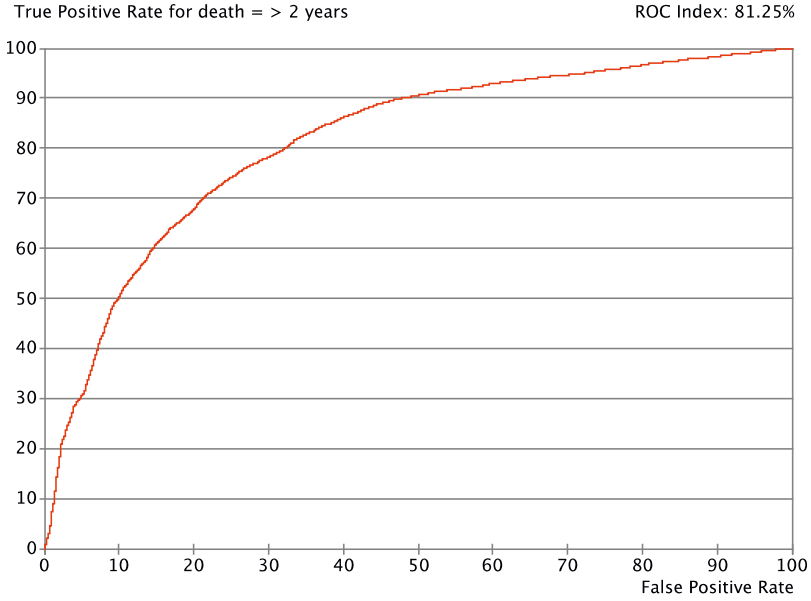


Figure 2: AUC for 2-year survival (0.81), computed using BayesiaLab and the test set comprising all editions

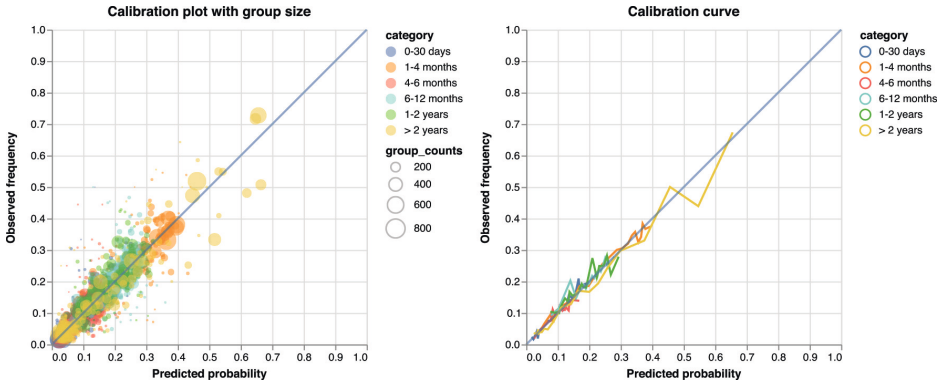


Figure 3: Visualization of model calibration for predicting survival. The calibration (bubble) plot shows, for each subpopulation, the relation between a survival category's predicted probability (x-axis) and its observed frequency in the dataset (y-axis). The size of each bubble corresponds to the population's size in the dataset. The calibration curve shows the same relation, but averages predicted/observed values by applying quantile-based binning with 1000 samples per bin (21 bins).

Tables

Period	TNM-edition	n (training)	n (training, after resampling)	n (test)
1999 – 2002:	TNM5	21528	45000	5323
2003 – 2009:	TNM6	43952	45000	11061
2010 – 2016:	TNM7	51378	45000	12842

Table 1: TNM-editions used by the Netherlands Cancer Registry and number of records available by period, split into data for training and testing.

Parameter	Values	Description
T	▪ T1 ▫ T1a ▫ T1b ▪ T2 ▫ T2a ▫ T2b ▪ T3 ▪ T4 ▪ TX	Clinical T descriptor (i.e. cT). The values T0 and Tis were excluded because of the low frequency in the cancer registry data.
N	▪ N0 ▪ N1 ▪ N2 ▪ N3	Clinical N descriptor (i.e. cN)
M	▪ M0 ▪ M1 ▫ M1a ▫ M1b	Clinical M descriptor (i.e. cM)

Parameter	Values	Description
TNM	▪ 1a ▪ 1b ▪ 2a ▪ 2b ▪ 3a ▪ 3b ▪ 4 ▪ X	Clinical stage group (i.e. cTNM).
incidence_year	Integer	Year of incidence
days_follow_up	Integer	Years of follow-up available
vital_status	Boolean	Vital status during last follow-up
survival	▪ < 30 days ▪ 1 – 4 months ▪ 4 – 6 months ▪ 6 – 12 months ▪ 1 – 2 years	Time of death after diagnosis. Calculated using days of follow and vital status.
multifocal	Boolean	True if the tumor was multifocal.
tumor size	Integer	Diameter in mm. Only available for patients with incidence year ≥ 2015.
sulcus superior	Boolean	Indicates involvement of the sulcus superior. Only available for patients with incidence year ≥ 2015.

Table 2: Parameters obtained from the Netherlands Cancer Registry (NCR). The variables below the thick line were only used for computing the test set and not for training the model.

	Predicting TNM6		Predicting TNM7		Predicting survival	
	macro	micro	macro	micro	macro	micro
sensitivity	0.918	0.957	0.851	0.948	0.282	0.350
specificity	0.994	0.994	0.993	0.993	0.862	0.870
accuracy	0.989	0.989	0.987	0.987	0.783	0.783

Table 3: macro- and micro aggregated statistics for predicting 6th edition stage group using 7th edition data, the 7th edition stage group using 6th edition data and survival using data from all editions.

References

1. Greene FL, Sobin LH. The staging of cancer: a retrospective and prospective appraisal. *CA Cancer J Clin*. 2008;58(3):180-190. doi:10.3322/CA.2008.0001
2. Amin MB, Edge S, Greene F, et al., eds. *AJCC Cancer Staging Manual*. 8th ed. Springer International Publishing; 2017.
3. Goldstraw P, Crowley J, Chansky K, et al. The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours. *J Thorac Oncol Off Publ Int Assoc Study Lung Cancer*. 2007;2(8):706-714. doi:10.1097/JTO.0b013e31812f3c1a
4. Edge SB, American Joint Committee on Cancer, eds. *AJCC Cancer Staging Manual*. 7th ed. New York: Springer; 2010.
5. Boffa DJ, Detterbeck FC, Smith EJ, et al. Should the 7th edition of the lung cancer stage classification system change treatment algorithms in non-small cell lung cancer? *J Thorac Oncol Off Publ Int Assoc Study Lung Cancer*. 2010;5(11):1779-1783. doi:10.1097/JTO.0b013e3181ee80c7
6. Pearl J. *Causality: Models, Reasoning and Inference*. 2nd ed. New York, NY, USA: Cambridge University Press; 2009.
7. Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques. MIT Press; 2009.
8. Conrady S, Jouffe L. Bayesian Networks & BayesiaLab - A Practical Introduction for Researchers.; 2015.
9. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J R Stat Soc Ser B Methodol*. 1977;39(1):1-38.
10. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag*. 2009;45:427-437. doi:10.1016/j.ipm.2009.03.002
11. Perez F, Granger BE. IPython: A System for Interactive Scientific Computing. *Comput Sci Eng*. 2007;9(3):21-29. doi:10.1109/MCSE.2007.53
12. McKinney W. Data Structures for Statistical Computing in Python. In: *Proceedings of the 9th Python in Science Conference*. ; 2010:51-56.
13. Oliphant TE. *Guide to NumPy*. 2nd ed. USA: CreateSpace Independent Publishing Platform; 2015.
14. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
15. Niculescu-Mizil A, Caruana R. Predicting Good Probabilities with Supervised Learning. In: *Proceedings of the 22Nd International Conference on Machine Learning*. ICML '05. New York, NY, USA: ACM; 2005:625–632. doi:10.1145/1102351.1102430

Supplementary Materials

In-depth analysis of frequent misclassifications

A number of misclassifications occur in stages II and III when using the BN with a MAP-query (which was originally used for prediction). Specifically, this happens for stages 2B and 3B when predicting the 6th edition and 2A and 2B when predicting the 7th edition. These groups are discussed more in-depth below.

7th to 6th edition misclassifications

Misclassified 2Bs

What should have been 2B in the 6th edition was frequently misclassified as 1B. Upon inspection of the data, it appears that all 2Bs that were misclassified as 1B were caused by T3N0M0 cases. For this subgroup, the BN (incorrectly) predicts a T3 in TNM7 to be a T2 (58% probability) instead of a T3 (35% probability) in TNM6.

TNM 7 (input)				**TNM 6** (actual)				**TNM 6** (predicted)			
T	**N**	**M**	**TNM**	**T**	**N**	**M**	**TNM**	**T**	**N**	**M**	**TNM**
3	0	0	2B	3	0	0	2B	2	0	0	1B
3	0	0	2B	2	0	0	1B	2	0	0	1B

Table S1: Relation between TNM predicted and actual values. The correct prediction is shown in grey.

The definitions of class T3 in the 6th and 7th edition largely overlap, but differ in a few key aspects, one of which affects this situation: the 7th edition shifts tumors ≥ 70mm from T2 to T3, solely based on tumor diameter. Therefore, a T3 ≥ 70mm in TNM7 may be a T2 in TNM6 if no other T3-criteria, such as invasion of the chest wall, apply.

Since we didn't have data on these other criteria, we were unable to definitively distinguish T2s from T3s during manual reclassification. Consequently, we cannot verify the validity of these probabilities.

Misclassified 3Bs

What should have been 3B in the 6th edition was frequently misclassified as 1B, 2B or 3A. This appears related to the fact that in the 6th edition, multifocal tumors limited to the same lobe would be considered T4, where in the 7th edition these tumors are classified as T3.

When inspecting the BN's behavior, the reason for misclassifying the 3B stage, appears to be the same as above: in all cases the BN misclassified a TNM7 T3 as T2 (probability 58%) instead of T4 (probability 4%).

TNM 7 (input)				**TNM 6** (actual)				**TNM 6** (predicted)			
T	**N**	**M**	**TNM**	**T**	**N**	**M**	**TNM**	**T**	**N**	**M**	**TNM**
3	0	0	2B	4	0	0	3B	2	0	0	1B
3	1	0	3A	4	1	0	3B	2	1	0	2B
3	2	0	3A	4	2	0	3B	2	2	0	3A

Table S2: Relation between TNM values in cases where 3B was misclassified as 1B, 2B or 3A.

6th to 7th edition misclassification

Misclassified 2Asa

What should have been 2A in the 7th edition was frequently misclassified as stage 1B. When looking at the data, this was caused by T2N0M0 cases.

In the 7th edition, the T2-class was split into T2A and T2B. In TNM7 a T2A N0 M0 would imply stage group 1B where T2B N0 M0 would imply stage group 2B. Here, the network misclassified a T2 as T2A (probability 58%) instead of T2B (probability 18%).

TNM 6 (input)				**TNM 7** (actual)				**TNM 7** (predicted)			
T	**N**	**M**	**TNM**	**T**	**N**	**M**	**TNM**	**T**	**N**	**M**	**TNM**
2	0	0	1B	2A	0	0	1B	2A	0	0	1B
2	0	0	1B	2B	0	0	2A	2A	0	0	1B

Table S3: Relation between TNM values in cases where 2A was misclassified as 1B. The correct prediction is shown in grey.

Misclassified 2Bs

What should have been 2B in the 7th edition was frequently misclassified as stage 2A or 3A.

When looking at the data, the first misclassification was caused by T2N1M0 cases; this is essentially the same situation as described above in “Misclassified 2As”.

The second issue is caused by T4M0M0 cases. Here the network considered a T4 in TNM6 to be a T4 in TNM7 instead of a T3. This is related to what was described above in “Misclassified 3Bs” (the fact that multifocal tumors limited to the same lobe are considered T3 in the 7th edition).

TNM 6 (input)				**TNM 7** (actual)				**TNM 7** (predicted)			
T	**N**	**M**	**TNM**	**T**	**N**	**M**	**TNM**	**T**	**N**	**M**	**TNM**
2	1	0	2B	2A	1	0	2A	2A	1	0	2A
2	1	0	2B	2B	1	0	2B	2A	1	0	2A
4	0	0	3B	3	0	0	2B	4	0	0	3A
4	0	0	3B	4	0	0	3A	4	0	0	3A

Table S4: Relation between TNM values in cases where 2B was misclassified as 2A or 3A. The correct prediction is shown in grey.

Conclusions

The mistakes seem to be related to situations where the definition of the T-descriptor has changed, causing cases to be shifted. However, considering the network is making predictions based on very limited information, this was to be expected.

Calculating micro- and macro-averages

In multi-class prediction, the outcome is not binary, but rather a class label. Evaluating the performance, for example using sensitivity or specificity, would therefore yield multiple metrics: one for each class. Macro- and micro-averaging are methods to reduce multiple metrics (e.g. multiple specificities) into a single metric. They differ in the way they deal with class imbalances: macro-averaging treats all classes equally (a class with only 10 samples has the same influence as a class that holds > 1000 samples) where in micro-averaging a class' influence on the final measure is weighted by its size. The formulas we used are listed below.

$$microsensitivity = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} TP_c + FN_c}$$

Equation S1: Computation of the micro-sensitivity. Here TP_c denotes the number of true positives for class c. Likewise, FN_c denotes the number of false negatives. C represents the total number of classes.

$$macrosensitivity = \sum_{c=1}^{C} \frac{TP_c / TP_c + FN_c}{C}$$

Equation S2: Macro-sensitivity. Here TP_c denotes the number of true positives for class c. Likewise, FN_c denotes the number of false negatives. C represents the total number of classes.

$$microspecificity = \frac{\sum_{c=1}^{C} TN_c}{\sum_{c=1}^{C} TN_c + FP_c}$$

Equation S3: Micro-specificity. Here TN_c denotes the number of true negatives for class c. Likewise, FP_c denotes the number of false positives. C represents the total number of classes.

$$macrospecificity = \sum_{c=1}^{C} \frac{TN_c / TN_c + FP_c}{C}$$

Equation S4: Macro-specificity. Here TN_c denotes the number of true negatives for class c. Likewise, FP_c denotes the number of false positives. C represents the total number of classes.

$$microaccuracy = macroaccuracy = \frac{\sum_{c=1}^{C} TP_c + TN_c}{\sum_{c=1}^{C} P_c + N_c}$$

Equation S5: Micro- and macro-accuracy. Here TP_c denotes the number of true negatives for class c. Likewise, TN_c denotes the number of true negatives. P_c and N_c represent the total number of positive and negative cases for a class (summed they make up the total number of cases in the dataset). C represents the total number of classes.

Supplemental Figures

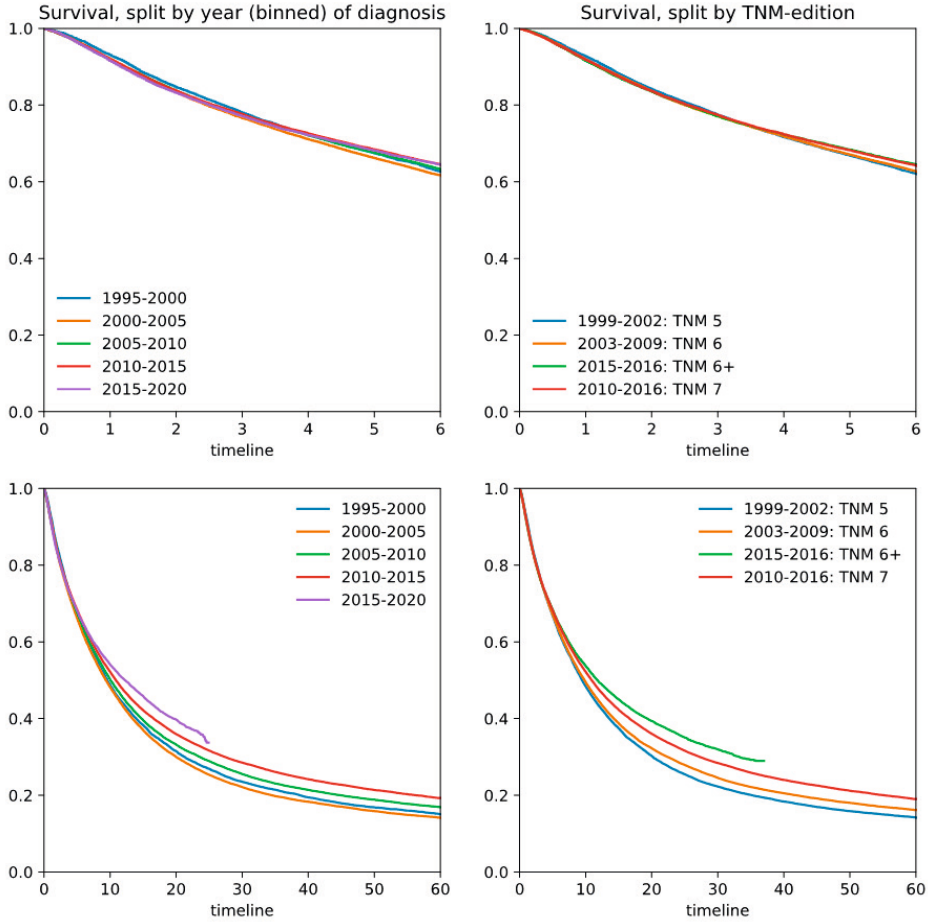


Figure S1: Kaplan-Meier plots with confidence intervals (timeline in months). The top panels are zoomed 10x with respect to the bottom panels. The left panels show survival stratified by (binned) year of diagnosis. The right panels show survival stratified according to the period for each TNM edition. Here, TNM6+ marks the timeframe that was manually reclassified from 7th to 6th edition. The blue vertical lines indicate 1, 4, 6, 12 and 24 months.

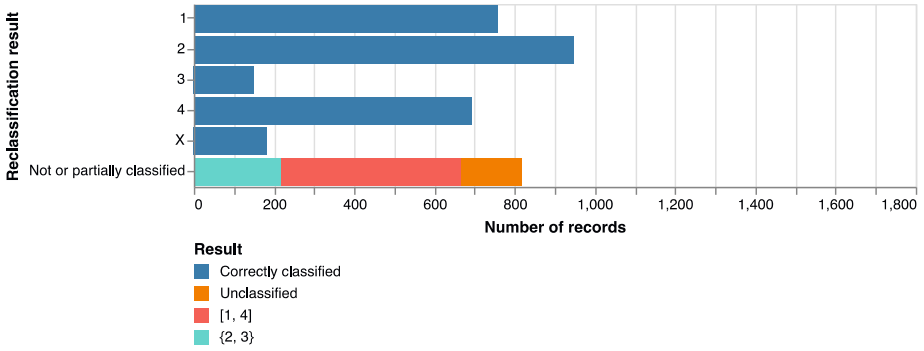


Figure S2: Distribution of TNM7 data reclassified as TNM 6 – *T-descriptor*. Partial classification results are denoted using brackets (e.g. [1, 4]) for ranges and accolades (e.g. {2, 3}) for sets.

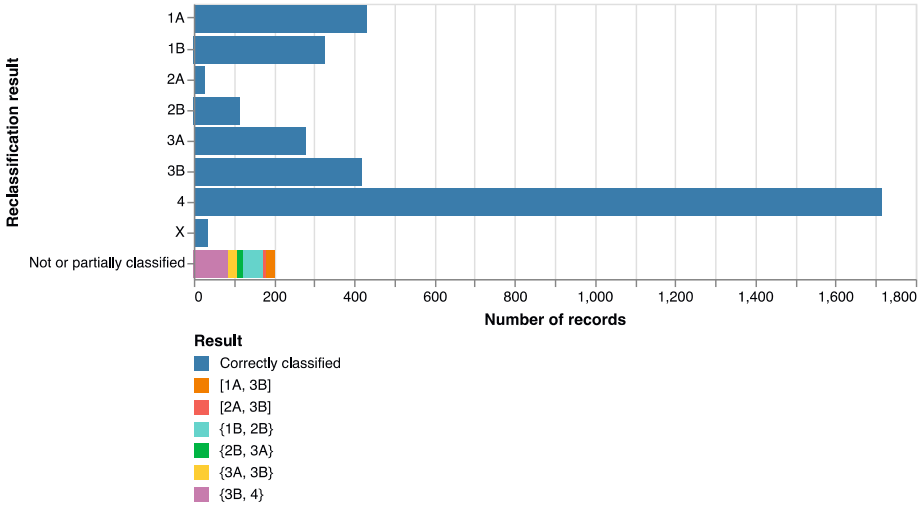


Figure S3: Distribution of TNM7 data reclassified as TNM6 – TNM-stage grouping. Partial classification results are denoted using brackets (e.g. [1A, 3B]) for ranges and accolades (e.g. {2B, 3A}) for sets.

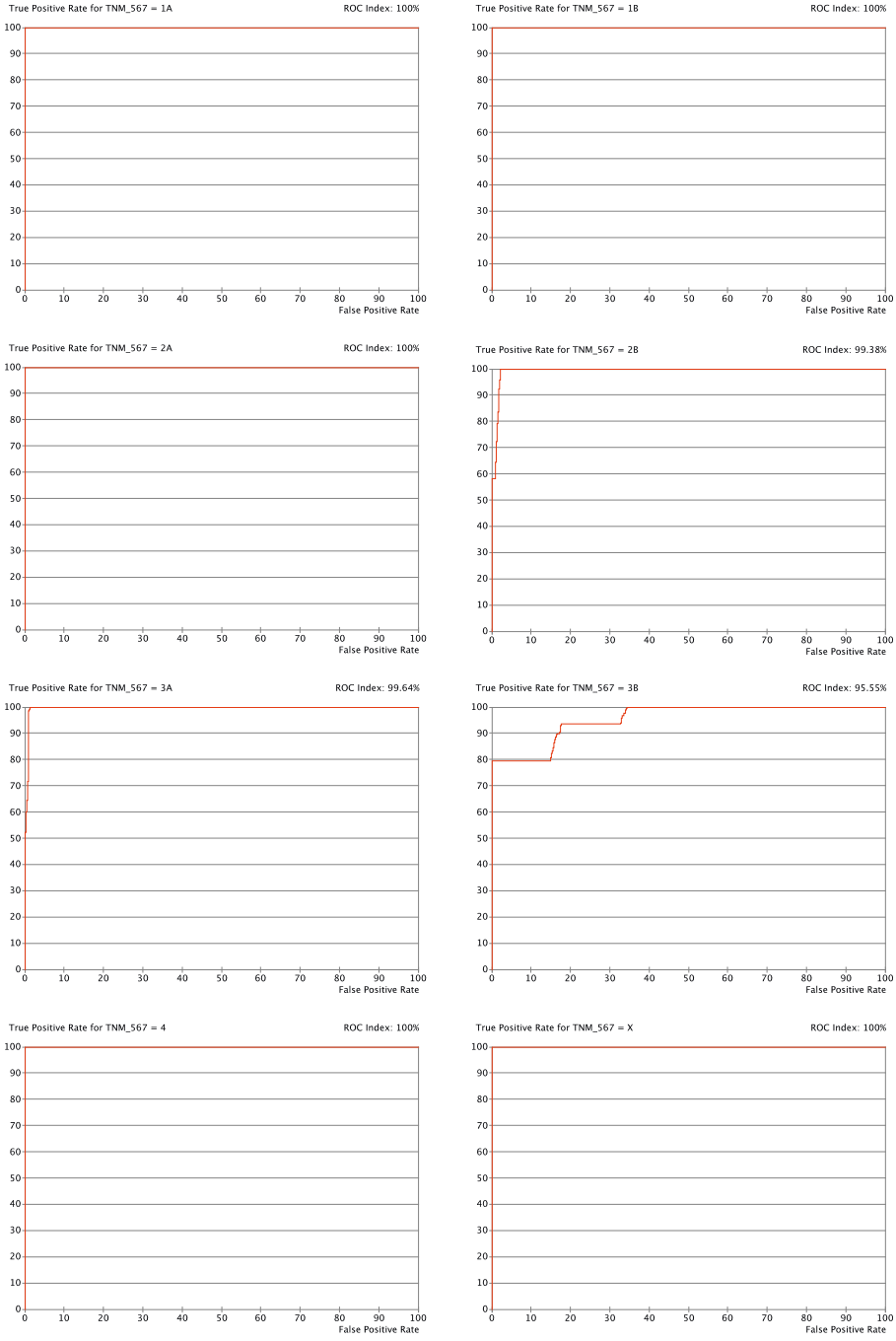


Figure S4: AUC curves for predicting 6th edition stage group using 7th edition data

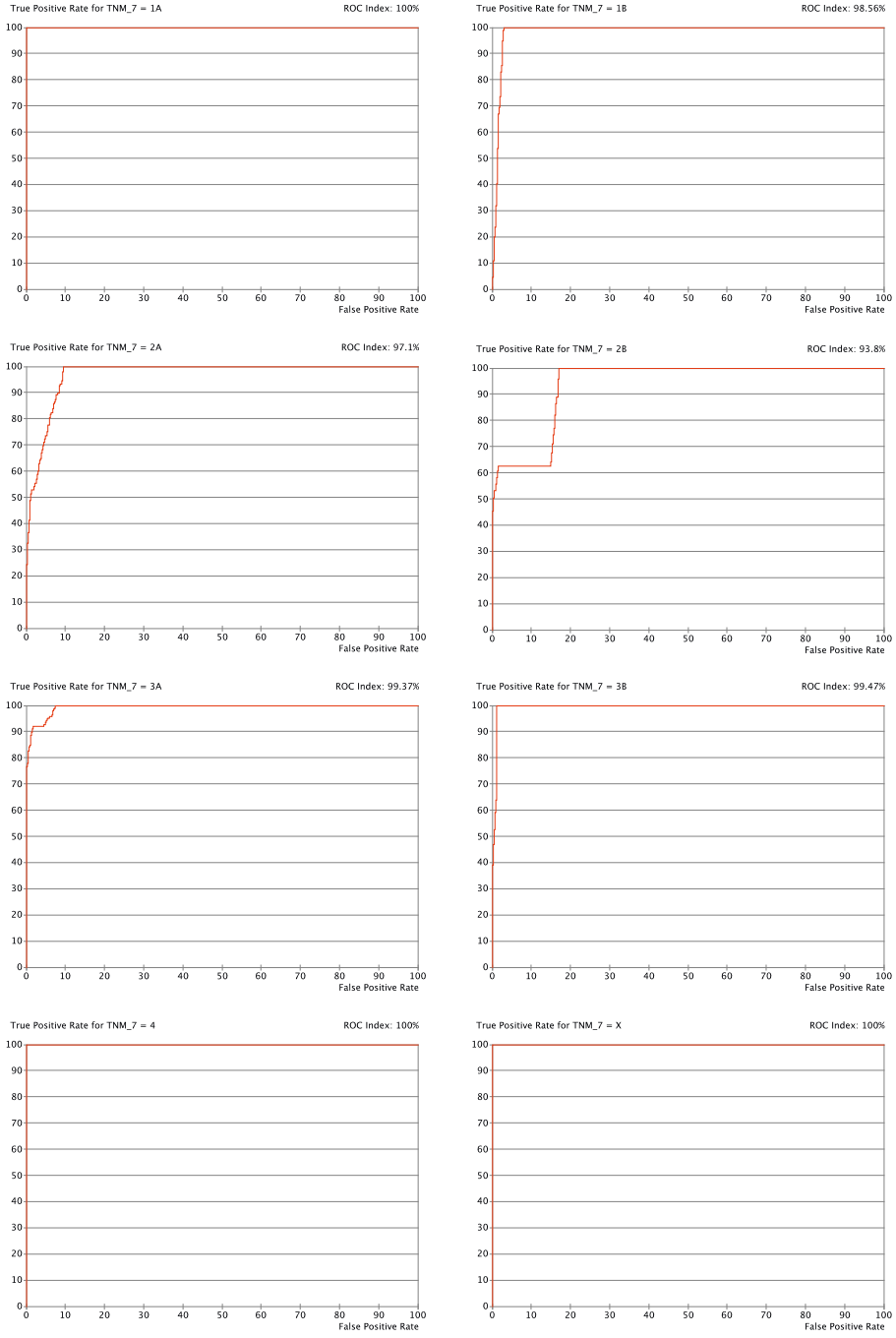


Figure S5: AUC curves for predicting 7^{th} edition stage group using 6^{th} edition data

Supplemental Tables

Edition	TNM 5								TNM 6								TNM 7							
T_7	1a	1b	2a	2b	3	4	X		1a	1b	2a	2b	3	4	X		1a	1b	2a	2b	3	4	X	
T_567																								
1	0.93	0.93	0.01	0.01	0.01	0.01	0.01		0.93	0.93	0.01	0.01	0.01	0.01	0.01		0.01	0.01	0.01	0.01	0.01	0.01	0.01	
1a	0.01	0.01	0.01	0.01	0.01	0.01	0.01		0.01	0.01	0.01	0.01	0.01	0.01	0.01		0.93	0.01	0.01	0.01	0.01	0.01	0.01	
1b	0.01	0.01	0.01	0.01	0.01	0.01	0.01		0.01	0.01	0.01	0.01	0.01	0.01	0.01		0.01	0.93	0.01	0.01	0.01	0.01	0.01	
2	0.01	0.01	0.93	0.93	0.47	0.01	0.01		0.01	0.01	0.93	0.93	0.47	0.01	0.01		0.01	0.01	0.01	0.01	0.01	0.01	0.01	
2a	0.01	0.01	0.01	0.01	0.01	0.01	0.01		0.01	0.01	0.01	0.01	0.01	0.01	0.01		0.01	0.01	0.93	0.01	0.01	0.01	0.01	
2b	0.01	0.01	0.01	0.01	0.01	0.01	0.01		0.01	0.01	0.01	0.01	0.01	0.01	0.01		0.01	0.01	0.01	0.93	0.01	0.01	0.01	
3	0.01	0.01	0.01	0.01	0.47	0.31	0.01		0.01	0.01	0.01	0.01	0.47	0.31	0.01		0.01	0.01	0.01	0.01	0.93	0.01	0.01	
4	0.01	0.01	0.01	0.01	0.01	0.62	0.01		0.01	0.01	0.01	0.01	0.01	0.62	0.01		0.01	0.01	0.01	0.01	0.01	0.93	0.01	
X	0.01	0.01	0.01	0.01	0.01	0.01	0.93		0.01	0.01	0.01	0.01	0.01	0.01	0.93		0.01	0.01	0.01	0.01	0.01	0.01	0.93	

Table S5: Conditional probabilities used as priors for node T_567, estimated using the AJCC staging manual. As an example, the highlighted cell shows the probability $P(T\_567 = 1|T\_7 = 1b, edition = TNM5)$.

Edition	TNM 5					TNM 6					TNM 7				
N_7	0	1	2	3	X	0	1	2	3	X	0	1	2	3	X
N_567															
0	0.96	0.01	0.01	0.01	0.01	0.96	0.01	0.01	0.01	0.01	0.96	0.01	0.01	0.01	0.01
1	0.01	0.96	0.01	0.01	0.01	0.01	0.96	0.01	0.01	0.01	0.01	0.96	0.01	0.01	0.01
2	0.01	0.01	0.96	0.01	0.01	0.01	0.01	0.96	0.01	0.01	0.01	0.01	0.96	0.01	0.01
3	0.01	0.01	0.01	0.96	0.01	0.01	0.01	0.01	0.96	0.01	0.01	0.01	0.01	0.96	0.01
X	0.01	0.01	0.01	0.01	0.96	0.01	0.01	0.01	0.01	0.96	0.01	0.01	0.01	0.01	0.96

Table S6: Conditional probabilities used as priors for node N_567, estimated using the AJCC staging manual.

Edition	TNM 5			TNM 6			TNM 7		
M_7	0	1a	1b	0	1a	1b	0	1a	1b
M_567									
0	0.97	0.01	0.01	0.97	0.01	0.01	0.97	0.01	0.01
1	0.01	0.97	0.97	0.01	0.97	0.97	0.01	0.01	0.01
1a	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.97	0.01
1b	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.97

Table S7: Conditional probabilities used as priors for node M_567, estimated using the AJCC staging manual.

TNM 7					TNM 6	
T	M	multifocal	tumor diameter	sulcus superior	T	M
1A	any	any	any	any	1	
1B	any	any	any	any	1	
2A	any	any	any	any	2	
2B	any	any	any	any	2	
3	any	no	< 70 mm	any	3	
3	any	no	≥ 70 mm	yes	3	
3	any	no	≥ 70 mm	no	2 – 3	
3	any	yes	any	any	4	
4	any	no	any	any	4	
4	0	yes	any	any	1 – 4	0 – 1
any	1A	any	any	any		1
any	1B	any	any	any		1

Table S8: rules for reclassifying TNM7 into TNM6. Rules should be applied top to bottom. Mapping the N-descriptor is omitted since it is a one-to-one mapping.

Actual Stage Group (TNM 6) → [predicted]↓	1A	1B	2A	2B	3A	3B	4	X
1A	431	0	0	0	0	0	0	0
1B	0	327	0	48	0	44	0	0
2A	0	0	27	0	0	0	0	0
2B	0	0	0	67	10	15	0	0
3A	0	0	0	0	268	27	0	0
3B	0	0	0	0	0	331	0	0
4	0	0	0	0	0	0	1715	0
X	0	0	0	0	0	0	0	34

Table S9: Confusion matrix for predicting 6th edition stage group using 7th edition data. The column headers show the actual stage, the row headers show the predicted stage.

Actual Stage Group (TNM 7) → [predicted]↓	1A	1B	2A	2B	3A	3B	4	X
1A	431	0	0	0	0	0	0	0
1B	0	248	79	0	0	0	0	0
2A	0	0	69	25	0	0	0	0
2B	0	0	0	48	0	0	0	0
3A	0	0	0	44	360	0	0	0
3B	0	0	0	0	27	264	0	0
4	0	0	0	0	0	0	1715	0
X	0	0	0	0	0	0	0	34

Table S10: Confusion matrix for predicting 7th edition stage group using 6th edition data. The column headers show the actual stage, the row headers show the predicted stage.

Actual survival → [predicted] ↓	0-30 days	1-4 months	4-6 months	6-12 months	1-2 years	> 2 years
0-30 days	2	0	1	3	0	0
1-4 months	1548	3596	1331	2228	1417	527
4-6 months	0	0	0	0	0	0
6-12 months	176	591	335	814	713	444
1-2 years	77	258	168	465	509	441
> 2 years	139	430	288	876	1248	2472

Table S11: Confusion matrix for predicting survival using data from 5th, 6th and 7th edition. The column headers show the actual survival, the row headers show the predicted survival.

	TNM 7 to 6			TNM 6 to 7		
	sensitivity	specificity	ROC-AUC	sensitivity	specificity	ROC-AUC
1A	1.00	1.00	1.00	1.00	1.00	1.00
1B	1.00	0.97	1.00	1.00	0.97	0.99
2A	1.00	1.00	1.00	0.47	0.99	0.97
2B	0.58	0.99	0.99	0.41	1.00	0.94
3A	0.96	0.99	0.96	0.93	0.99	0.99
3B	0.79	1.00	0.96	1.00	0.99	0.99
4	1.00	1.00	1.00	1.00	1.00	1.00
X	1.00	1.00	1.00	1.00	1.00	1.00

Table S12: Sensitivity, specificity and ROC estimates for predicting each stage group.

4

Prognostic Factors Analysis for Oral Cavity Cancer Survival in the Netherlands and Taiwan using a Privacy-Preserving Federated Infrastructure

Adapted from Geleijnse G., Chiang R.C.J., Sieswerda M., Schuurman M., Lee K.C., van Soest J., Dekker A., Lee W.C., Verbeek X.A.A.M.; "**Prognostic factors analysis for oral cavity cancer survival in the Netherlands and Taiwan using a privacy-preserving federated infrastructure**"; *Sci Rep.* 2020;10(1):20526. doi:10.1038/s41598-020-77476-2

Abstract

Background: The difference in incidence of oral cavity cancer (OCC) between Taiwan and the Netherlands is striking. Different risk factors and treatment expertise may result in survival differences between the two countries. However, due to regulatory restrictions, patient-level analyses of combined data from the Netherlands and Taiwan are infeasible.

Methods: We implemented a software infrastructure for federated analyses on data from multiple organizations. We included 41,633 patients with single-tumor OCC between 2004 and 2016, undergoing surgery, from the Taiwan Cancer Registry and Netherlands Cancer Registry. Federated Cox Proportional Hazard was used to analyze associations between patient and tumor characteristics, country, treatment and hospital volume with survival.

Results: Five factors showed differential effects on survival of OCC patients in the Netherlands and Taiwan: age at diagnosis, stage, grade, treatment and hospital volume. The risk of death for OCC patients younger than 60 years, with advanced stage, higher grade or receiving adjuvant therapy after surgery was lower in the Netherlands than in Taiwan; but patients older than 70 years, with early stage, lower grade and receiving surgery alone in the Netherlands were at higher risk of death than those in Taiwan. The mortality risk of OCC in Taiwanese patients treated in hospitals with higher hospital volume (≥50 surgeries per year) was lower than in Dutch patients.

Conclusions: We conducted analyses without exchanging patient-level information, overcoming barriers for sharing privacy sensitive information. The outcomes of patients treated in the Netherlands and Taiwan were slightly different after controlling for other prognostic factors.

Introduction

The difference in incidence of oral cavity cancer (OCC) between both the Netherlands and Taiwan is striking. Taiwan has one of the world's highest incidence rates of OCC [1]. In 2016, 5,116 patients were diagnosed with OCC with a standardized incidence rate of 13.8 cases per 100,000 population. Men in Taiwan are at 10.8 times higher risk for OCC than women [2]. Contrary to Taiwan, OCC in the Netherlands is a rare disease with an annual incidence of approximately 900 cases (or 5.5/100,000 inhabitants) [3]. Changes in incidence, mortality and survival may reflect changes in risk factors, diagnostics, clinicopathological factors, and treatment [4,5]. To be able to provide high standards of care, the treatment of head and neck tumours in the Netherlands is centralized within 14 expertise centres. Expertise and patients' characteristics may result in survival differences between these different geographical areas. Also, prognostic factors for OCC survival may have differential effects in patients of these two countries. However due to regulatory restrictions, patient-level analyses where data is shared between these countries is unfeasible.

With the implementation of the General Data Protection Regulation (GDPR) in the European Union, cancer registries are amidst an on-going debate on its implications.[6] The GDPR may be one of the arguments for data processing entities such as cancer registries to be reticent in data sharing initiatives. In particular, the GDPR poses restrictions on data sharing with parties outside the European economic area, including Taiwan. In collaborations such as Eurocare, Globocan and RARECARE, cancer registries are sharing patient-level data to facilitate large scale international epidemiological research [7–9]. In such international studies, the patient record data are typically delivered to a trusted organization, responsible for processing the pooled data. Said regulations and privacy concerns pose a threat to the continuation of these initiatives.

Innovations in information technology have created an alternative to the traditional pooling of data. Ohno-Machado and colleagues developed a series of algorithms "building shared models without sharing data", in order to compute regression models without record level data leaving the participating organizations [10–12]. Feasibility of federated privacy-preserving classification algorithms and survival analyses have been demonstrated using mathematical and experimental analyses [13–16]. Several other machine learning algorithms were created for "distributed learning" and successfully applied it to a number of studies in involving multiple radiotherapy centres using a commercial software application [17,18]. We developed an open source implementation of a federated privacy preserving data analysis platform [19,20]. Unlike other initiatives [21–23], it offers a flexible open-source infrastructure that allows to deploy federated algorithms implemented in wide range of programming languages. Hence, the infrastructure allows to deploy existing algorithms as described in literature and combine them into a series of analyses. Also, it does not assume a prescribed data format, which makes it suitable for cancer registries.

In this work, we apply the federated privacy preserving data analysis platform to compare the prognostic factors for OCC survival between Taiwan and the Netherlands.

Methods

Data

The Taiwan Cancer Registry (TCR) is a national population-based cancer registry system established in 1979. Data from Taiwanese patients with newly diagnosed malignancies in hospitals with 50 or more beds are mandatory reported to the TCR. Details of the history, objectives, and activities of the TCR have been well-documented [24]. With its high data quality and completeness (approximately 98%), the TCR is also one of the highest-quality cancer registries in the world [25].

The Netherlands Cancer Registry (NCR) is a nationwide registry in which all newly diagnosed malignancies in the Netherlands are documented. It has a nationwide coverage since 1989. The main source of notification of the NCR is the automated nationwide network and registry of histo- and cytopathology (PALGA) and it is complemented by other sources such as the National Registry of Discharge Diagnosis. After notification, specially trained registry clerks routinely extract data on patients and tumour characteristics from patient's medical records in all Dutch hospitals.

The dataset for each registry follows standard research protocols, and the selected variables are converted into defined code. All patients who underwent surgery for a diagnosis with an oral cavity squamous cell carcinoma (ICD-O Topography codes C00.3-5, C02, C03, C04, C05.0, C05.8, C05.9, C06 and morphology codes 8050-8089) between 2004 and 2016 were selected from the TCR and NCR. In case of multiple primary OCCs, only the first primary tumour was included in the study. In both registries, tumour topography, morphology and grade were coded according to the International Classification for Disease Oncology 3rd Edition-(ICD-O-3). For tumour stage, the Netherlands uses the Tumour Node Metastases (UICC TNM, 6th and 7th editions), whereas Taiwan adopts the AJCC 6th and 7th editions cancer staging system. However, the UICC and AJCC cancer staging systems are almost the same, so the staging data are comparable. Treatment was categorized in primary surgery and surgery with adjuvant radiotherapy and/or chemotherapy. Hospital volume was defined as the number of OCC surgeries performed in the centre where the patient was treated in the year of the patient's diagnosis. Volume was divided into 3 categories (<50, 50-99 and ≥100 surgeries/year). Survival was defined as the time from date of diagnosis to date of death or until the last date of follow-up. Data on vital status and date of death through linkage with the population death databases were collected up to January 31, 2019. This study was approved by Netherlands Cancer Registry's Supervisory Committee (K18.098) and the National Taiwan University Hospital Research Ethnics Committee (201801116RINA).

Federated Infrastructure

To enable the privacy-preserving analysis of the Dutch and Taiwanese data, open source software was written to facilitate the analysis of local data and communication of aggregated statistics. We created a software infrastructure, where a server coordinates computing tasks and is connected via the internet to the computers (nodes) of the two organizations. The system conceptually consists of three components: a central server, multiple nodes and (software run by) a researcher. Each participating site runs a node that has access to the patient-level data and connects to the central server. The central server handles administrative tasks like authentication and authorization, and acts as a central point for communication between the nodes. Software run by researchers can upload "tasks", for example "compute the sums over all columns", to the central server, which are picked up by the nodes and executed. While tasks run on patient level data, the nodes only return aggregated data, no patient identifiable data is shared. Multiple tasks can be chained to create a script including more complex or iterative algorithms. Orchestration is then performed by software run on the researcher's computer. A more detailed and technical description of the infrastructure as well as all open source software can be found at the website [19].

Statistical Analysis

The means or frequencies of patient characteristics, treatment modalities and hospital volume were compared between countries. Chi-square test was used for analysing categorical variables. A federated version of the Cox proportional hazard algorithm with Breslow's method for ties was implemented [19]. Mathematical decomposition of the algorithm and its soundness were demonstrated by Lu and colleagues [12]. Briefly, the nodes iteratively compute aggregated statistics based on the latest estimates of the Hazard Ratios (HRs) and the local registry data. Next, the aggregated statistics from the sites are combined to compute an updated estimation of the HRs. Finally, the estimation of the HRs has converged, the algorithm finishes. We also performed interaction analyses to assess whether the prognostic factors of OCC are different or have differential effects on survival between the Netherlands and Taiwan. P-values for interaction are based on the likelihood ratio test of the interaction term between "country" and the respective prognostic factors.

Following Lu et al, we implemented the Newton-Raphson update to iteratively estimate the HRs for the selected covariates. This implementation is known to converge quickly (i.e. require few iterations), but it requires complex computations for each iteration. To restrict the complexity, we use a follow up period in years with one decimal rather than a period in days. In our analysis, the algorithm terminates when the difference between the sums of the previous and updated HRs after an iteration is less than 10^{-8}.

Results

Patient characteristics

A total of 7,766 and 33,867 newly diagnosed OCC cases with single primary tumor and receiving surgical treatment were recorded from 2004 to 2016 in the Netherlands and Taiwan, respectively (Table 1). In the Netherlands, the mean age was 63.9 years and among them, 44% were men. However, the mean age in Taiwan was 10 years younger (53.3 years) than the Netherlands and more than 91% of patients were men. The common sites of OCC in the Netherlands were floor of the mouth and gum (41.8%) and other/unspecified parts of tongue (41.2%); but in Taiwan, the common sites were buccal and other parts of mouth (44.8%) and other/unspecified parts of tongue (36.3%). Additionally, most patients in the Netherlands were treated in hospitals with the lowest hospital volume (<50 oral surgery/year, 59%), while in Taiwan, nearly two-third of patients received treatment in hospitals with the highest hospital volume (≥100 oral surgery/year, 64%). Similarly, period of diagnosis, cancer stage, tumor grade, and treatment modalities between the two countries were all significantly different.

Univariable analyses

In Table 2, the univariable cox regression model for Dutch data and Taiwanese data is performed separately by each country, whereas the combined data is analysed at each site using the privacy-preserving federated algorithm. Our findings showed that increasing age, male gender, higher stage, poorer differentiation grade, surgery with adjuvant radiation and/or chemotherapy, and location (e.g. floor of mouth, gum, buccal, and other parts of mouth) were all significant prognostic factors for shorter survival in both the Netherlands and Taiwan. However, period of diagnosis and hospital volume are influential prognostic factors for longer survival in Taiwan, but not in the Netherlands. In combined data, without adjusting for other factors, OCC patients in the Netherlands had worse overall survival than those in Taiwan (HR, 1.39; 95% CI 1.34-1.44). Additionally, the hazard ratio pattern of all prognostic factors, except gender, is similar between individual data and combined data. With regard to gender, Dutch data and Taiwan data show that women's overall survival rate is significantly better than men's; however, in the combined data, because of the higher survival rate of male patients in Taiwan, the survival curve of women crosses the curve of men. Therefore, the HR of gender in the combined data shows no significance (HR, 0.97; 95% CI 0.93-1.01).

Multivariable analyses

As shown in Table 3, younger age at diagnosis, female gender, recent years at diagnosis, early stage, well differentiated grade, receiving primary surgery alone, and higher hospital volume were all significant independent prognostic factors for longer survival in the combined data. After adjusting for other prognostic factors, including age, gender, period of diagnosis, stage, location, grade, treatment, and hospital volume, patients with OCC in Taiwan had slightly better outcomes than those in the Netherlands (HR, 1.06; 95%

CI 1.01-1.12). Moreover, only patients with hard palate cancers (HR, 1.30; 95% CI 1.17-1.45) had poorer survival after adjusting other covariables. Patients with surgery and adjuvant radiation and/or chemotherapy (HR, 1.40; 95% CI 1.34-1.46) had poorer survival than those with primary surgery alone. Compared with patients treated in the hospitals with ≥100 oral cavity surgeries/year, patients treated in the hospitals with <50 surgeries/year (HR 1.13; 95% CI 1.08-1.08) were independently associated with a poorer survival.

Prognostic factors with significant factor-by-country interaction are shown in Figure 1; the following factors have differential effects on survival of OCC patients in the Netherlands and Taiwan: age at diagnosis, stage, tumour grade, treatment, and hospital volume. First, the mortality risk of OCC patients in the Netherlands and Taiwan both increased with increasing age; however, Dutch patients had a stronger association between risk of dying and increasing age than Taiwanese patients. The risk of death for patients younger than 60 years was slightly lower in the Netherlands than in Taiwan, but patients older than 70 years in the Netherlands were at higher risk of death than those in Taiwan. Second, higher stage increased the risk of death of OCC patients in both the Netherlands and Taiwan. However, the risk increments are different in the two countries such that early staged OCC patients had higher risk of death in the Netherlands than those in Taiwan, whereas patients with advanced stage in the Netherlands were at lower risk of death than in Taiwan. Third, the risk of death for patients with well and moderately differentiated grade was slightly lower in Taiwan than in the Netherlands. However, patients with poorly differentiated grade in Taiwan were at higher risk of death than those in the Netherlands. Fourth, OCC patients receiving surgery alone in the Netherlands had much higher risk of death than those in Taiwan; but the risk of death for patients in Taiwan receiving adjuvant radiotherapy and/or chemotherapy was higher than those in the Netherlands. And finally, the mortality risk of OCC in Taiwanese patients who were treated in hospitals with higher hospital volume (≥50 surgeries per year) was lower than in Dutch patients. However, patients treated in hospitals with lower hospital volume (<50 surgeries per year) had similar outcomes in both the Netherlands and Taiwan.

Discussion

The aetiology of OCC in the Netherlands and Taiwan are different. Although smoking and alcohol consumption are the major risk factors in both countries, betel nut chewing is an important risk factor for Taiwan, which may explain why the incidence rates differ greatly between the two countries. Patient characteristics and experts' experience with treating this disease may result in survival differences between the Netherlands and Taiwan. Our findings also confirmed the prognostic factors of oral cancer reported in previous studies [26–28]. In the present study, we found that the outcomes of patients treated in the Netherlands and Taiwan were slightly different after controlling for other prognostic factors. As for the potential prognostic factors, we found that age at diagnosis,

gender, period of diagnosis, stage, tumour grade, treatment modalities, and hospital volume significantly influence the survival of OCC patients. Five prognostic factors (age, stage, grade, treatment modality and hospital volume) of OCC have differential effects on survival between the Netherlands and Taiwan.

As mentioned previously, we found that the mortality risk of OCC patients, both in the Netherlands and Taiwan, increased with increasing age, higher stage and poorer differentiation grade after adjusting for other prognostic factors in both countries (Figure 1). Older patients in the Netherlands are at higher risk of death than in Taiwan. Elderly patients may have multiple comorbidities that affect the choice of treatment and tolerance to treatment; therefore, the burden of comorbidity among older OCC patients may be larger in the Netherlands than in Taiwan. Patients with advanced stage and poorly differentiated grade in Taiwan have a higher risk of death than those in the Netherlands. The presence of extranodal extension may be related to the severity of tumour stage and grade, thereby it may influence the prognosis differences in Taiwan and the Netherlands.[29] Meanwhile, extranodal extension in metastatic lymph nodes is an important predictor of regional recurrence and distant metastasis, and it is related to the poor prognosis of OCC [29]. Data on extranodal extension has been collected in Taiwan since 2011 and in the Netherlands since 2015. Additional information on co-morbid conditions and extranodal extension should be considered in future studies.

Surgery alone is the first choice for OCC treatment both in the Netherlands and Taiwan. Surgeons' experiences, such as complete resection with a tumour-free margin and comprehensive neck dissection, may be critical points in OCC treatment and prognosis. Previous studies in Taiwan showed that patients treated in hospitals with high surgery volume had better OCC survival [30,31]. The risk of death for OCC patients receiving surgery alone in the Netherlands was much higher than in Taiwan, which may be due to differences in surgical experience and/or patient selection. Additionally, previous studies showed that Dutch patients treated in hospitals with different volumes did not differ significantly; this may be due to its highly centralized treatment of head and neck tumours in the Netherlands [32]. However, the mortality risk of OCC in Taiwanese patients who were treated in hospitals with higher hospital volume (≥50 surgeries per year) was lower than in Dutch patients; thereby there may still be opportunities for improvement of OCC care in the Netherlands.

Nowadays, the clinical guidelines state that postoperative chemoradiation is recommended for patients with extranodal extension, but is also considered as an alternative to adjuvant radiotherapy for patients with positive surgical margins, pT3 to T4 primary tumours, pN2 to pN3 lymph node disease, perineural invasion, and lymphovascular invasion to improve control rates [33]. Although the risk of death for patients in Taiwan receiving adjuvant radiotherapy and/or chemotherapy was higher than those in the Netherlands, this difference might be explained by unmeasured

pathological characteristics, such as resection margins status, extranodal extension, perineural invasion, lymphovascular invasion, performance status and comorbidity. The lack of this information is a limitation of our research. Otherwise, the removal of known risk factors including smoking and alcohol drinking even after diagnosis may reduce the risk of recurrences and second tumours in existing oral cancer patients and also improve the prognosis [34]. However, it is less clear to date how the delays in diagnosis or treatment affect the cancer stage at diagnosis and survival in oral cancer patients [35]. Therefore, these relevant factors, such as lack of individual life-style habits and delays in diagnosis or treatment, should be considered in future research.

Today, only a limited amount of analyses has been developed for the federated infrastructure. For routine use, however, the infrastructure needs to be extended with all commonly used algorithms for data analyses. The main limitation of this work is that the algorithm to check the proportional hazard assumption was not yet implemented. The multivariable regression (18 covariates, Table 3) required a computation time of around 6 minutes, and the coefficients converged after 5 iterations. Alternative implementations may better deal with higher dimensional data [36]. By design, visual inspection of tables with patient data is not supported. Accordingly, performing federated data analysis will require a different way of working. Advanced quality checking software and adding algorithms for descriptive statistics may mitigate this limitation, as they allow to better understand quality and limitations of datasets [37].

In the past, combined and interaction analyses of individual patient data from different countries required sharing data between different parties and processing the pooled data in a designated central lab. As it respects patient privacy and complies to data protection regulations, the federated analysis of distributed data platform makes it possible to perform analyses of individual patient data without exchanging patient-level information. To enable this study, we successfully developed an open source IT infrastructure that allows the deployment of algorithms for federated analysis of distributed data and used it for survival analysis for OCCs on patient data from the Netherlands and Taiwan Cancer Registries. This work is the first application of this technology to enable analyses of data from multiple cancer registries. In future work, this infrastructure can be expanded with exploratory analyses and other regression and classification algorithms. Moreover, it can be applied to train artificial intelligence models on multimodal data, including imaging [38,39]. For studies where individual datasets are insufficient (e.g. in international comparisons and studies on rare cancers), the use of a federated infrastructure may become the de-facto standard.

Tables & Figures

	NETHERLANDS		TAIWAN		
	Cases	%	Cases	%	P value
Total	7766	100.0	33867	100.0	
Age (average)	63.9		53.3		
< 60 years	2709	34.9	24493	72.3	<.001
60 – 69 years	2542	32.7	6196	18.3	
≥ 70 years	2515	32.4	3178	9.4	
Gender					
Male	4356	44.0	30913	91.3	<.001
Female	3410	56.0	2954	8.7	
Period of diagnosis					
2004 – 2007	2148	30.2	7873	23.2	<.001
2008 – 2011	2400	33.6	10528	31.1	
2012 – 2016	3218	36.2	15466	45.7	
Stage					
I	3392	43.8	11239	33.2	<.001
II	1220	15.8	6918	20.4	
III	827	10.8	3946	11.7	
IVA	2208	28.1	10269	30.3	
IVB	64	0.8	969	2.9	
IVC	17	0.2	81	0.2	
Unknown	38	0.5	445	1.3	
Early stage	4612	59.4	18157	53.6	<.001
Advanced stage	3116	40.1	15265	45.1	
Unknown	38	0.5	445	1.3	
Location					
Mucosa of lip (ICD-O C003-005)	114	1.5	728	2.1	<.001
Other/unspecified parts of tongue (ICD-O C02)	3215	41.2	12282	36.3	
Floor of mouth and gum (ICD-O C03-04)	3234	41.8	5049	14.9	
Hard palate (ICD-O C050, C058-059)	117	1.4	636	1.9	
Buccal and other parts of mouth (ICD-O C06)	1086	14.1	15172	44.8	
Grade					
Well differentiated	1183	15.1	11285	33.3	<.001
Moderately differentiated	4084	52.1	17677	52.2	
Poorly or undifferentiated	1075	14.3	2355	7.0	
Unknown	1424	18.5	2550	7.5	

	NETHERLANDS		TAIWAN		
	Cases	%	Cases	%	P value
Treatment					
Primary surgery	4876	63.0	18570	54.8	<.001
Surgery with radiotherapy and/or chemotherapy	2890	37.0	15297	45.2	
Hospital volume (oral cavity surgeries/year)					
< 50	4466	59.3	5272	15.6	<.001
50 – 99	2560	34.7	6992	20.6	
≥ 100	740	6.0	21603	63.8	

Table 1: Patient characteristics.

	NETHERLANDS		TAIWAN		COMBINED	
	HR	95% CI	HR	95% CI	HR	95% CI
Country						
Taiwan	–	–	–	–	1.00	–
The Netherlands	–	–	–	–	1.39	1.34 – 1.44
Age						
< 60 years	1.00	–	1.00	–	1.00	–
60 – 69 years	1.48	1.36 – 1.62	1.19	1.13 – 1.24	1.27	1.22 – 1.32
≥ 70 years	2.55	2.36 – 2.77	1.93	1.83 – 2.03	2.16	2.07 – 2.24
Gender						
Female	1.00	–	1.00	–	1.00	–
Male	1.18	1.10 – 1.26	1.11	1.04 – 1.19	0.97	0.93 – 1.01
Period of diagnosis						
2004 – 2007	1.00	–	1.00	–	1.00	–
2008 – 2011	0.93	0.86 – 1.01	0.84	0.81 – 0.88	0.85	0.82 – 0.89
2012 – 2016	0.93	0.85 – 1.01	0.72	0.68 – 0.75	0.75	0.72 – 0.78
Stage						
Early stage	1.00	–	1.00	–	1.00	–
Advanced stage	2.12	1.99 – 2.26	3.11	2.99 – 3.23	2.78	2.69 – 2.88
Unknown	1.34	0.85 – 2.11	1.65	1.41 – 1.92	1.45	1.26 – 1.67
Location						
Other/unspecified parts of tongue	1.00	–	1.00	–	1.00	–
Mucosa of lip	0.81	0.60 – 1.09	0.99	0.87 – 1.13	0.95	0.84 – 1.07
Floor of mouth and gum	1.34	1.25 – 1.44	1.36	1.29 – 1.43	1.42	1.36 – 1.48
Hard palate	1.18	0.89 – 1.56	1.79	1.60 – 2.00	1.65	1.49 – 1.83
Buccal and other parts of mouth	1.29	1.17 – 1.43	1.07	1.03 – 1.12	1.04	1.01 – 1.08
Grade						
Well differentiated	1.00	–	1.00	–	1.00	–
Moderately differentiated	1.64	1.48 – 1.83	1.51	1.45 – 1.58	1.57	1.51 – 1.63
Poorly or undifferentiated	2.28	2.01 – 2.58	2.65	2.49 – 2.83	2.63	2.49 – 2.78
Unknown	1.49	1.32 – 1.69	1.02	0.94 – 1.10	1.23	1.16 – 1.31
Treatment						
Primary surgery	1.00	–	1.00	–	1.00	–
Surgery with radiotherapy and/or chemotherapy	1.62	1.52 – 1.73	2.60	2.50 – 2.70	2.29	2.22 – 2.36
Hospital volume (oral cavity surgeries/year)						
≥ 100	1.00	–	1.00	–	1.00	–
50 – 99	0.94	0.82 – 1.07	1.05	0.99 – 1.09	1.13	1.08 – 1.17
< 50	0.91	0.81 – 1.04	1.11	1.05 – 1.16	1.24	1.20 – 1.29

Table 2: Univariable Cox Regression Analyses. The figures in the Netherlands and Taiwan columns are computed locally, while the Combined column was computed using the privacy-preserving federated algorithm.

	NETHERLANDS		TAIWAN		COMBINED	
	HR	95% CI	HR	95% CI	HR	95% CI
Country						
Taiwan	–	–	–	–	1.00	–
The Netherlands	–	–	–	–	1.06	1.01 – 1.12
Age						
< 60 years	1.00	–	1.00	–	1.00	–
60 –69 years	1.49	1.36 – 1.62	1.29	1.23 – 1.35	1.31	1.26 – 1.36
≥ 70 years	2.80	2.58 – 3.04	2.30	2.18 – 2.43	2.40	2.30 – 2.50
Gender						
Female	1.00	–	1.00	–	1.00	–
Male	1.27	1.19 – 1.36	1.23	1.15 – 1.32	1.23	1.18 – 1.29
Period of diagnosis						
2004 –2007	1.00	–	1.00	–	1.00	–
2008 –2011	0.93	0.86 – 1.01	0.82	0.78 – 0.86	0.85	0.82 – 0.88
2012 –2016	0.87	0.80 – 0.96	0.67	0.64 – 0.70	0.73	0.70 – 0.76
Stage						
Early stage	1.00	–	1.00	–	1.00	–
Advanced stage	1.93	1.78 – 2.10	2.29	2.18 – 2.41	2.19	2.10 – 2.29
Unknown	1.46	0.93 – 2.31	1.57	1.35 – 1.83	1.46	1.26 – 1.69
Location						
Other/unspecified parts of tongue	1.00	–	1.00	–	1.00	–
Mucosa of lip	0.76	0.56 – 1.02	1.13	0.99 – 1.28	1.06	0.94 – 1.19
Floor of mouth and gum	1.11	1.03 – 1.20	0.98	0.92 – 1.03	1.01	0.97 – 1.06
Hard palate	0.92	0.70 – 1.22	1.40	1.25 – 1.57	1.30	1.17 – 1.45
Buccal and other parts of mouth	1.02	0.92 – 1.13	1.01	0.97 – 1.05	1.00	0.96 – 1.04
Grade						
Well differentiated	1.00	–	1.00	–	1.00	–
Moderately differentiated	1.51	1.35 – 1.68	1.36	1.31 – 1.42	1.38	1.33 – 1.44
Poorly or undifferentiated	1.90	1.67 – 2.15	2.06	1.93 – 2.20	1.92	1.82 – 2.03
Unknown	1.50	1.33 – 1.70	1.10	1.02 – 1.19	1.24	1.17 – 1.32
Treatment						
Primary surgery	1.00	–	1.00	–	1.00	–
Surgery and radiotherapy and/or chemotherapy	1.04	0.96 – 1.13	1.52	1.45 – 1.60	1.40	1.34 – 1.46
Hospital volume (oral cavity surgeries/year)						
≥ 100	1.00	–	1.00	–	1.00	–
50 – 99	0.90	0.79 – 1.03	1.00	0.95 – 1.04	1.01	0.97 – 1.05
< 50	0.94	0.82 – 1.07	1.19	1.13 – 1.25	1.13	1.08 – 1.18

Table 3: Multivariable Cox Regression Analyses. The figures in the Netherlands and Taiwan columns are computed locally, while the Combined column was computed using the privacy-preserving federated algorithm.

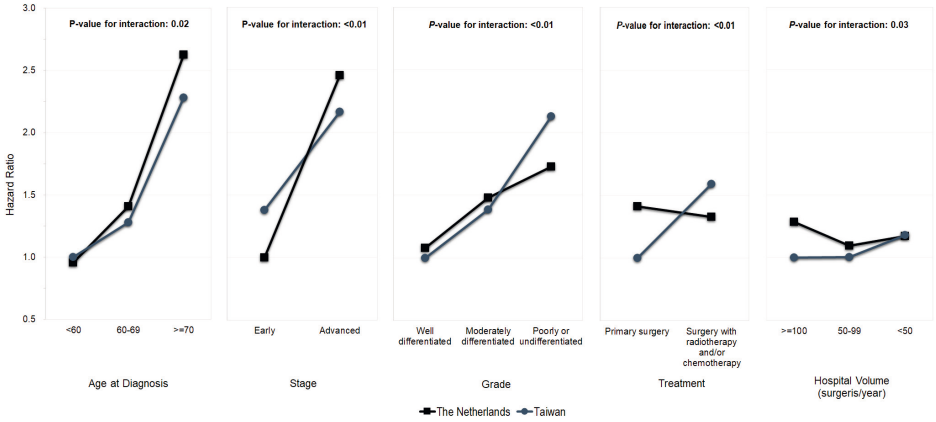
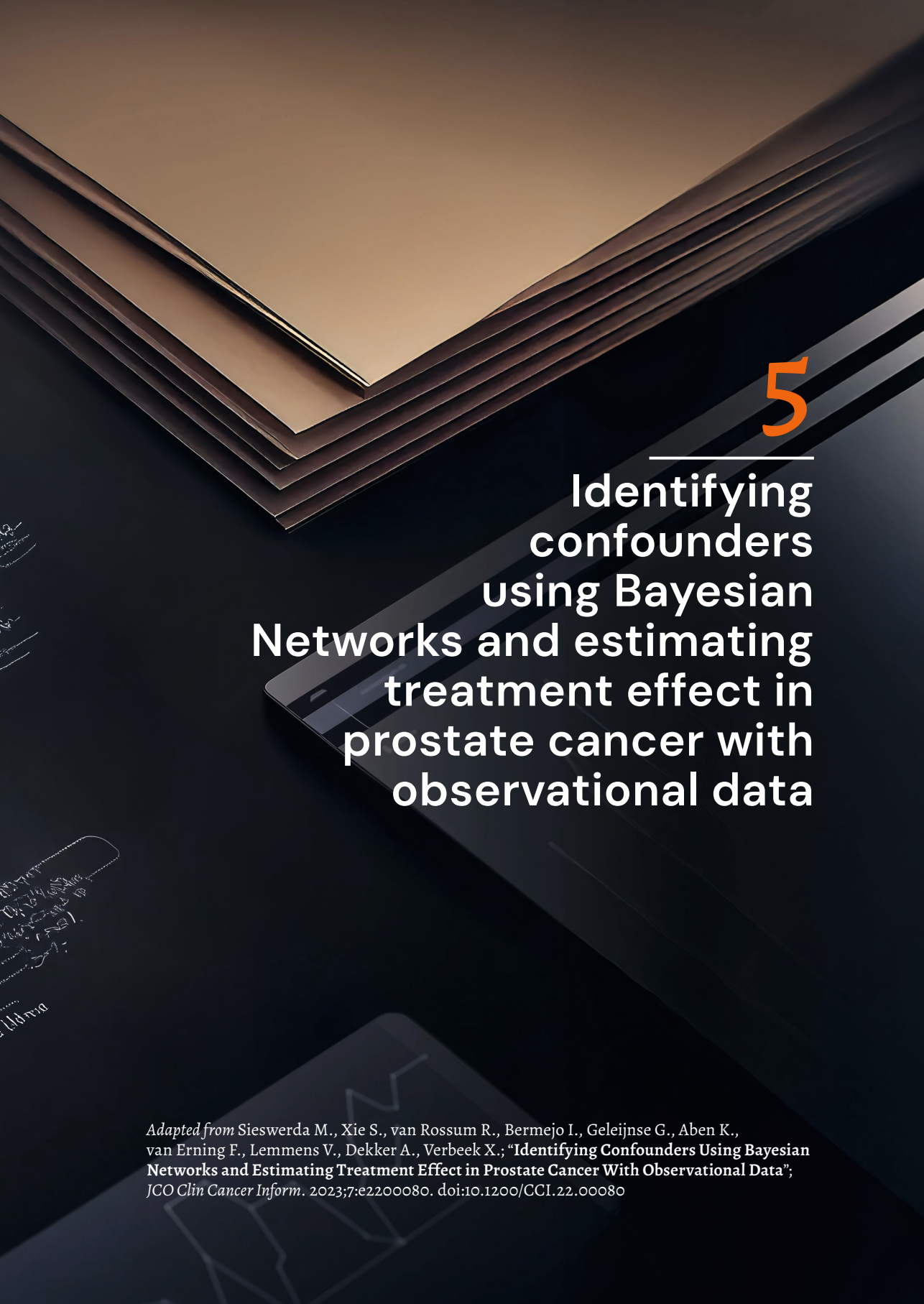


Figure 1: Interaction effects between country and five other prognostic factors.

References

1. Hsu W.-L., Yu K. J., Chiang C.-J., Chen T.-C. & Wang C.-P. Head and Neck Cancer Incidence Trends in Taiwan, 1980 2014. *International Journal of Head and Neck Science* **1**, 180–190 (2017).
2. Taiwan Cancer Registry Annual Report 2016. (2016).
3. The Netherlands Cancer Registry. https://www.cijfersoverkanker.nl/?language=en (2019).
4. van Dijk, B. A. C., Brands, M. T., Geurts, S. M. E., Merkx, M. A. W. & Roodenburg, J. L. N. Trends in oral cavity cancer incidence, mortality, survival and treatment in the Netherlands: OCC Incidence, Mortality, Survival and Treatment. *Int. J. Cancer* **139**, 574–583 (2016).
5. Liu, S.-Y. *et al.* Surgical outcomes and prognostic factors of oral cancer associated with betel quid chewing and tobacco smoking in Taiwan. *Oral Oncology* **46**, 276–282 (2010).
6. van Veen, E.-B. Observational health research in Europe: understanding the General Data Protection Regulation and underlying debate. *European Journal of Cancer* **104**, 70–80 (2018).
7. Sant, M. *et al.* Cancer survival in Europe, 1999–2007: Doing better, feeling worse? *European Journal of Cancer* **51**, 2101–2103 (2015).
8. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012: Globocan 2012. *International Journal of Cancer* **136**, E359–E386 (2015).
9. Gatta, G. *et al.* Rare cancers are not so rare: The rare cancer burden in Europe. *European Journal of Cancer* **47**, 2493–2511 (2011).
10. Ohno-Machado, L. *et al.* iDASH: integrating data for analysis, anonymization, and sharing. *Journal of the American Medical Informatics Association* **19**, 196–201 (2012).
11. Wu, Y., Jiang, X., Kim, J. & Ohno-Machado, L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. *Journal of the American Medical Informatics Association* **19**, 758–764 (2012).
12. Lu, C.-L. *et al.* WebDISCO: A web service for distributed cox model learning without patient-level data sharing. *Journal of the American Medical Informatics Association* ocv083 (2015) doi:10.1093/jamia/ocv083.
13. Park, M. & Welling, M. A note on privacy preserving iteratively reweighted least squares. *arXiv:1605.07511 [cs, stat]* (2016).
14. Nguyên, T. T. & Hui, S. C. Privacy-Preserving Mechanisms for Parametric Survival Analysis with Weibull Distribution. *arXiv:1708.04517 [cs]* (2017).
15. McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & Arcas, B. A. y. Communication-Efficient Learning of Deep Networks from Decentralized Data. *arXiv:1602.05629 [cs]* (2016).
16. Zerka, F. *et al.* Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care. *JCO Clinical Cancer Informatics* 184–200 (2020) doi:10.1200/CCI.19.00047.
17. Jochems, A. *et al.* Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept. *Radiotherapy and Oncology* **121**, 459–467 (2016).
18. Deist, T. M. *et al.* Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clinical and Translational Radiation Oncology* **4**, 24–31 (2017).
19. IKNL. Vantage6.ai – Privacy preserving federated learning. https://www.vantage6.ai (2019).
20. Moncada-Torres, A., Martin, F., Sieswerda, M., van Soest, J. & Geleijnse, G. VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange. in *AMIA Annual Symposium Proceedings* (2020).
21. Tian, Y. *et al.* POPCORN: A web service for individual PrognOsis prediction based on multi-center clinical data CollabORatioN without patient-level data sharing. *Journal of Biomedical Informatics* **86**, 1–14 (2018).

22. Ryffel, T. *et al.* A generic framework for privacy preserving deep learning. *arXiv:1811.04017 [cs, stat]* (2018).
23. Jones, E. M. *et al.* DataSHIELD – shared individual-level analysis without sharing the data: a biostatistical perspective. *Norsk Epidemiologi* **21**, (2012).
24. Chiang, C.-J. *et al.* Quality assessment and improvement of nationwide cancer registration system in Taiwan: a review. *Japanese Journal of Clinical Oncology* **45**, 291–296 (2015).
25. Chiang, C.-J., Wang, Y.-W. & Lee, W.-C. Taiwan's Nationwide Cancer Registry System of 40 years: Past, present, and future. *Journal of the Formosan Medical Association* (2019) doi:10.1016/j.jfma.2019.01.012.
26. Gatta, G. *et al.* Prognoses and improvement for head and neck cancers diagnosed in Europe in early 2000s: The EUROCARE-5 population-based study. *European Journal of Cancer* **51**, 2130–2143 (2015).
27. Al-Swiahb, J. N. *et al.* Clinical, pathological and molecular determinants in squamous cell carcinoma of the oral cavity. *Future Oncology* **6**, 837–850 (2010).
28. Ghani, W. M. N. *et al.* Survival of Oral Cancer Patients in Different Ethnicities. *Cancer Investigation* **37**, 275–287 (2019).
29. Fang, K.-H. *et al.* Histological Differentiation of Primary Oral Squamous Cell Carcinomas in an Area of Betel Quid Chewing Prevalence. *Otolaryngol Head Neck Surg* **141**, 743–749 (2009).
30. Lin, C.-C. & Lin, H.-C. Effects of surgeon and hospital volume on 5-year survival rates following oral cancer resections: The experience of an Asian country. *Surgery* **143**, 343–351 (2008).
31. Chang, C.-M. *et al.* Multivariate Analyses to Assess the Effects of Surgeon and Hospital Volume on Cancer Survival Rates: A Nationwide Population-Based Study in Taiwan. *PLoS ONE* **7**, e40590 (2012).
32. de Ridder, M. *et al.* Variation in head and neck cancer care in the Netherlands. *European Journal of Surgical Oncology (EJSO)* **43**, 1494–1502 (2017).
33. Chen, M. M. *et al.* Trends and variations in the use of adjuvant therapy for patients with head and neck cancer: Adjuvant Therapy for Head and Neck Cancer. *Cancer* **120**, 3353–3360 (2014).
34. Warnakulasuriya, S. Living with oral cancer: Epidemiology with particular reference to prevalence and life-style changes that influence survival. *Oral Oncology* **46**, 407–410 (2010).
35. Gigliotti, J., Madathil, S. & Makhoul, N. Delays in oral cavity cancer. *International Journal of Oral and Maxillofacial Surgery* **48**, 1131–1137 (2019).
36. Conn, A. R., Scheinberg, K. & Vicente, L. N. *Introduction to derivative-free optimization*. (SIAM, Soc. for Industrial and Applied Math. [u.a.], 2009).
37. Martos, C. & Emanuele Crocetti. A proposal on cancer data quality checks: one common procedure for European cancer registries – version 1.1,. (2018).
38. Ilhan, B., Lin, K., Guneri, P. & Wilder-Smith, P. Improving Oral Cancer Outcomes with Imaging and Artificial Intelligence. *J Dent Res* **99**, 241–248 (2020).
39. Ariji, Y. *et al.* Contrast-enhanced computed tomography image assessment of cervical lymph node metastasis in patients with oral cancer by using a deep learning system of artificial intelligence. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology* **127**, 458–463 (2019).



5

Identifying confounders using Bayesian Networks and estimating treatment effect in prostate cancer with observational data

Adapted from Sieswerda M., Xie S., van Rossum R., Bermejo I., Geleijnse G., Aben K., van Erning F., Lemmens V., Dekker A., Verbeek X.; "**Identifying Confounders Using Bayesian Networks and Estimating Treatment Effect in Prostate Cancer With Observational Data**"; *JCO Clin Cancer Inform*. 2023;7:e2200080. doi:10.1200/CCI.22.00080

Abstract

PURPOSE: Randomized controlled trials are considered the golden standard for estimating treatment effect but are costly to perform and not always possible. Observational data, although readily available, is sensitive to biases such as confounding by indication. Structure learning algorithms for Bayesian Networks (BNs) can be used to discover the underlying model from data. This enables identification of confounders through graph analysis, although the model might contain noncausal edges. We propose using a blacklist to aid structure learning in finding causal relationships. This is illustrated by an analysis into the effect of active treatment (v observation) in localized prostate cancer.

METHODS: In total, 4,121 prostate cancer records were obtained from the Netherlands Cancer Registry. Subsequently, we developed a (causal) BN using structure learning while precluding noncausal relations. Additionally, we created several Cox proportional hazards models, each correcting for a different set of potential confounders (including propensity scores). Model predictions for overall survival were compared with expected survival on the basis of the general population using data from Statistics Netherlands (Centraal Bureau voor de Statistiek).

RESULTS: Structure learning precluding noncausal relations resulted in a causal graph but did not identify significant edges toward treatment; they were added manually. Graph analysis identified year of diagnosis and age as confounders. The BN predicted a treatment effect of 1 percentage point at 10 years. Chi-squared analysis found significant associations between year of diagnosis, age, stage, and treatment. Propensity score correction was successful. Adjusted Cox models predicted significant treatment effect around 3 percentage points at 10 years.

CONCLUSION: A blacklist in conjunction with structure learning can result in a causal BN that can be used for confounder identification. Treatment effect found here is close to the 5 percentage point found in the literature.

Introduction

When evaluating treatment effect, data from a double-blinded, randomized controlled trial (RCT) is considered the golden standard; randomization ensures that baseline characteristics between intervention and control group are similar. Unfortunately, RCTs have limitations too. They are costly to perform and tend to have short-term primary outcomes. Since they usually target a specific subpopulation, obtaining a sufficiently large study size can be difficult and results are frequently not generalizable. Furthermore, they often investigate a single intervention and are not always possible due to scientific or ethical concerns.

In contrast, in oncology, observational data is readily available, may cover the entire population and offer follow-up of many years. For example, the Netherlands Cancer Registry (NCR) encompasses a large body of data, representative of the population by design. Traditionally this data is used to answer questions about cancer incidence and prevalence, organization and quality of care, trends in treatment, and treatment effectiveness.[1–4]

However, using observational data to compare treatment effectiveness remains challenging due to selection biases. Specifically, confounding by indication may lead to over- or underestimation of treatment effect.[5–7] Here, we define a confounder as any variable with a causal effect on both treatment selection and outcome (see also Supplemental section Confounding). For example, when a curative treatment option is more frequently selected for relatively healthy patients, measured effect will be skewed towards better survival: prior health status acts as a confounder. Many techniques have been developed to address these biases. Even then, the question remains which variables to use for deconfounding.[8, 9]

The most common approaches are to select all pre-treatment variables (**pre-treatment criterion**) or all variables that are a common cause of treatment *and* outcome (**common cause criterion**). However, it can be shown that there are circumstances where these selections are insufficient and may even introduce bias (see Supplemental Figure S1).[10, 11] For example, in 2008, Giordano et al. contrasted results from analyses of data from the Surveillance, Epidemiology, and End Results (SEER) Program with knowledge obtained through RCTs.[5] In one analysis, where they compared effect of active treatment with observation in localized prostate cancer, they concluded that (propensity score) correction did not remove confounding; not only did the cancer population have a significantly lower other-cause mortality, but patients that underwent radical prostatectomy had a better prognosis than a control population without cancer.

Fortunately, this bias can sometimes be avoided. When the full causal structure is known, blocking all backdoor paths from treatment to outcome is sufficient to find a deconfounding set if it exists (**backdoor path criterion**).[12, 13] Unfortunately, the structure of the underlying model is usually unknown. Another method requires only that it is

known whether each (pre-treatment) variable is a cause for treatment and/or outcome (**disjunctive cause criterion**).[10] However, even this knowledge is not always available. Interestingly, algorithms are available to elucidate relationships between variables from data, for example for Bayesian Networks.

Bayesian Networks (BNs) are a type of probabilistic graphical model whose structure is determined by a directed acyclic graph (DAG) where nodes represent variables and directed edges signify (preferably causal) relationships.[13,14] Each node is associated with a probability distribution conditional on its parents (i.e., nodes with a directed edge towards that node). Consequently, conditional (in)dependencies can be read of the graph. Here we denote that X is conditionally independent of Z given Y as $X \perp\!\!\!\perp Z \mid Y$. Similarly, if X is conditionally dependent of Z given Y this is denoted as $X \not\perp\!\!\!\perp Z \mid Y$.

Previously, Häggström demonstrated that structure learning can be used to reproduce an underlying model from data which was subsequently used to identify a set of deconfounding variables, all without making assumptions about associations present in the data.[11] However, this approach ignores edge directions and thus the notion of causality, making results difficult to interpret.

Where conditional dependency is present in the data, a V-structure is induced (see Supplemental Figure S2). However, in other situations directionality cannot be determined (see also section Graph equivalence in the supplemental materials). This is further complicated where confounders are suspected: the networks $X \rightarrow Y, Y \leftarrow Z \rightarrow X$ and $Y \leftarrow X \rightarrow Z, Z \rightarrow Y$ represent the same probability distribution $P(X, Y, Z)$. Using data alone it is impossible to establish if Z is a mediator or confounder. Consequently, graphs found by plain structure learning are unlikely causal.

While in epidemiology causal relationships are frequently suspected, they are often the target of research and certainty about these relationships is rare. Counterfactuals, on the other hand, can often be supplied with reasonable certainty. For example, tumor characteristics, treatment, or outcome will never *cause* age even if the *reverse* relation is contended.

Interestingly, structure learning can be constrained using a blacklist with prohibited edge directions. Therefore, we hypothesize a blacklist can aid structure learning in finding causal relationships, by precluding directed edges known to be non-causal. This can be considered an extension of the method previously developed by Häggström.

Here we demonstrate this process and show how the result can be used for confounder identification and mitigation. Specifically, we investigate the effect of active treatment vs observation in localized prostate cancer, attempting to reproduce the previously mentioned analysis by Giordano et. al.[5] using data from the Netherlands Cancer Registry (NCR).

Methods

Data

Where possible, data selection and processing was aligned with the original analysis by Giordano et al.[5] Records of patients with prostate cancer, diagnosed between 2005 and 2014, aged between 66 and 80, with a cT1-2 tumor were retrieved from the NCR. While not explicitly described by Giordano et al., we additionally assumed cN0 and cM0. (See also Supplemental section "Breakdown of applying in-/exclusion criteria")

Active treatment was defined as primary radiotherapy or prostatectomy (See also Supplemental section "Determining active treatment"). 646 cases where treatment started later than 180 days after diagnosis were classified as observation. Patients with primary hormonal therapy, chemotherapy, targeted therapy, or radiotherapy on metastases were excluded, as were patients that died within 1 year of diagnosis.

For assessing association with treatment and for use in the BN, age was discretized into the categories [66–69], [70–74] and [75–80]. A new variable stage ("tumor size" in Giordano's manuscript), was calculated with values I (cT ≤ 2A) and II (cT ≥ 2B). Grade was derived from the (pre-treatment) Gleason score, with grade 1 (well differentiated) corresponding to Gleason score 2-4, grade 2 (moderately differentiated) to 5-6 and grade 3 (poorly differentiated) corresponding to 7-10. Records with grade 3 were excluded as well as those where grade could not be calculated due to missing values.[15] The level of prostate-specific antigen (PSA) was discretized into the categories low (< 10 ng/mL), elevated (10 ≤ PSA < 20 ng/mL), and high (> 20 ng/mL). Income & education could be represented by socioeconomic status (SES) in deciles as estimated by the Netherlands Institute for Social Research using partial zip code. SES was discretized into three categories (high: 10-8, medium: 7-4, and low: 3-1). The variables race, urban residence, marital status, SEER region and comorbidity from the original analysis were not available in any capacity.

Overall survival, counted from initial diagnosis, was discretized into Boolean variables representing 1- to 10-year survival, introducing NAs for patients that had a follow-up of less than 10 years and were alive at the time of follow-up.

The final dataset consisted of 4121 patient records.

Propensity score analysis

Again, processing was aligned with the original analysis by Giordano et al.[5] The Chi^2-test was used to assess the association between variables and treatment (Table 1). Propensity scores were calculated using logistic regression (on treatment) with covariables age, SES, year of diagnosis, stage and grade and subsequently, discretized using quintiles. The Cochran-Mantel-Haenszel (CMH) Chi^2-test, with the bins as strata, was used to evaluate the effect of propensity score correction.[16–18]

Bayesian network development

A BN was developed using Hugin, version 7.4.[19] Structure learning and parameter estimation were performed using the NPC-algorithm (with Level of Significance set to 0.05) and the EM-algorithm respectively.[20, 21]

The NPC-algorithm may yield ambiguous regions, consisting of "a set of inter-dependent uncertain links" where the absence of one link depends on the presence of another. When this happens, Hugin prompts for input. For resolution we made associations with treatment or survival take precedence. Further analysis was completed using the Python package Thomas.[22]

When applying structure learning without constraints, Hugin identified one ambiguous region involving year of diagnosis, age and 10y-surival, which was resolved by selecting year of diagnosis - 10y-surival (Supplemental Figure S3).

Next, based on clinical knowledge, a blacklist was defined. Specifically, edges known to be non-causal were precluded (Table 2; see also Supplemental section "Bayesian network development"). Using the blacklist, Hugin identified the same ambiguous region as before. However, upon manual resolution, Hugin ignored the blacklist and directed the edge from 10y-surival towards year of diagnosis (Supplemental Figure S4). The correctly directed edge was added to a minimal whitelist and structure learning was repeated. This led to another ambiguous region and Hugin misdirecting surv_09y à age, which was handled similarly. These steps are shown in Figure 1. Finally, several edges were manually added to facilitate survival analysis as defined in the *extended* whitelist after which the process was repeated, and network parameters were estimated (Supplemental Figure S5). A flow-diagram of the development process is shown in Supplemental Figure S6.

Variables confounding treatment were identified through inspection of the graph and treatment effect was estimated. First, by calculating $P(survival \mid treatment)$, considering treatment an observation. Then, through causal reasoning and computing $P(survival \mid do(treatment))$, considering treatment an intervention.[13, 14]

Structural differences between all four graphs were quantified using the Structural Hamming Distance (SHD).[23]

Cox Proportional Hazards Models

To evaluate the effect of including and excluding variables in regression analyses, three "regular" Cox Proportional Hazards models and one "stratified" model were created, matching variables used in each of the models by Giordano et. al as close as possible (Table 3). Additionally, we created a model that fulfilled the backdoor criterion. Hazard ratios (HR) for treatment were extracted. Survival probabilities were calculated using covariates' statistical mode as central values.

Expected survival based on the general population

Probabilities for 1-to-10-year-survival in the general population, stratified by age, gender and incidence year, were downloaded from Statistics Netherlands (CBS) and matched with the NCR-data.[24] Mean expected-survival-probability and 95% confidence interval were calculated for each combination of treatment and age category.

Results

Variable distribution & propensity score correction

Chi^2 analysis showed significant association between treatment and age, stage, PSA, and year of diagnosis. Comparison with the original analysis was not possible, because these data were not published. All associations except PSA, which was not included in the propensity score calculation, resolved after stratification using the discretized propensity score as strata in the CMH-Chi^2-test.

Bayesian network development

The network obtained through structure learning without providing prior knowledge contained several non-causal edges (Figure 2, top left panel). These were, in addition to all edges between the survival nodes, grade → year of diagnosis, PSA → age, PSA → year of diagnosis, treatment → age, and 10y-surival → age. No associations between treatment and survival were found. The SHD between the graph without constraints and with minimal whitelist was 9 (Supplemental Table S1).

Applying structure learning while precluding non-causal edges also did not identify significant associations between treatment and survival; these only appeared when increasing the significance threshold to ~0.20. Additional edges between treatment and survival nodes were added as defined in the extended whitelist (Table 2). EM-learning was successfully applied. The final causal network is shown in the bottom left panel in Figure 2.

Conditional probabilities of survival given treatment were calculated using the BN. This suggested a (small) positive association of 4.24 percentage points between active treatment and 10-year survival (Supplemental Table S2).

Analysis of the final network structure identified year of diagnosis and age as confounders. Two additional backdoor paths existed through PSA. Either correcting for year of diagnosis and age, or deleting edges pointing towards treatment was sufficient to fulfil the backdoor criterion; a mutilated graph was created that excluded these edges.

The probabilities of survival given *do(treatment)* were calculated using the mutilated graph. This suggested a (small) *positive* effect of 1 percentage point of active treatment on survival (Supplemental Table S2).

Cox Proportional Hazards Models

Treatment was significantly associated with survival (at $p \leq 0.05$) in all "regular" models trying to match Giordano's (models 1 – 3), in two strata (q3 and q5) of model 4, and in the additional analysis that only included the confounders identified through the Bayesian Network (model 5). For each model, the HR for treatment and probability of 5- and 10-year survival are shown in Table 3.

Expected survival based on the general population

Mean expected-survival-probability and its 95% confidence interval were calculated for the three age groups and treatment options. Results were visualized and overlayed with predicted probabilities from the BNs and Cox-models as shown in Figure 3. All Cox-models predicted survival within the expected-survival confidence interval.

Discussion & conclusion

Directionality of edges can sometimes be induced through conditional dependencies estimated from data. However, this is usually not the case. In our analysis too, structure learning without prior knowledge resulted in a model that was clearly non-causal: treatment should never cause pre-treatment variables. Yet, by guiding structure learning with prior knowledge, a causal graph was obtained. The SHD of 9 edges shows that using a blacklist also impacts the modelled probability distribution. Structure learning did not find an association between treatment and survival at the selected significance threshold of 0.05; associations only appeared when increasing this threshold to ~0.20. This could be the result of the combination of (relatively) small sample size and limited treatment effect. To facilitate further analyses, edges were manually added. Once a causal graph was available, either the disjunctive cause or backdoor path criterion could be applied.

Opting for the disjunctive cause criterion, the BN suggested we should correct for age, stage, year of diagnosis and PSA, the same variables identified through Chi^2 analysis. According to the backdoor path criterion, we would only need to correct for year of diagnosis and age. The first criterion minimizes the risk of missing a confounder, the second the standard error of the estimated effect.

For example, structure learning did not identify PSA as an independent cause of survival while Chi^2 analysis, did show an association (data not shown). Assuming the BN is correct, this suggests the effect of PSA on survival is mediated through age and year of diagnosis and adjusting for it is unnecessary; treatment seems less likely as mediator since structure learning did not associate it with survival. Still, correcting for these

additional variables should not significantly affect the outcome; these would be bias neutral adjustments, although they may increase the standard error of the estimated effect.[25, 26] Of course, the possibility exists that edges between PSA or stage and survival were not found due to insufficient (statistical) power. If so, some bias remains uncorrected for when using the backdoor criterion, although one could argue the effects are limited, otherwise the edges would have been found.

All Cox-models found significant associations between treatment and survival, except for three strata in model 4. Adjusting for potential confounders (models 2, 3 and 5) increased the HR and *P*-value. As expected, correction based on the backdoor path criterion (model 5) resulted in a similar increase in HR as when including additional variables (models 2 and 3).

Quantitatively, the BN estimated the effect of treatment on 10y-survival at 1 percentage-point. Adjusted Cox-models suggested a 3 percentage-point survival benefit (HR 0.88 [0.79-0.99]). These results are in line with results from RCTs. Specifically, an RCT (n=695) reported by Bill-Axelson, Holmberg et al. found a relative risk of 0.74 (0.56 to 0.99), corresponding to a statistically significant reduction of 5 percentage points in overall survival at 10 years.[27] Still, the clinical relevance, in terms of gained overall survival, appears limited. Limited treatment effect could also explain why Hamdy et al. report a larger RCT (n= 1643) that did not find a statistically significant effect of active treatment in a younger, screen-detected population.[28]

However, our estimates are smaller than previous estimates based on SEER-data.[5, 15] Specifically, Wong et al. estimated an HR of 0.69 [0.66-0.72] when adjusting for grade, size and number of comorbidities and Giordano an HR of 0.68 [0.65-0.70] adjusting for age, stage, grade, SES, and propensity score. (See also Supplemental Table S3)

Since we could not assess disease-specific survival directly, we compared overall survival with expected survival based on the general population. All Cox-models adjusting for (potential) confounders predicted survival after treatment to be within the expected range for the general population. This, again, appears different from the results that Giordano et al. obtained, where other-cause-mortality was significantly lower in the treated population, even in their propensity corrected model. This might be explained by differences in screening practice, which is much more common in the United States (and potentially associated with higher education levels and SES); in contrast, in the Netherlands screening is cautiously recommended against.

We have shown how structure learning, guided by clinical knowledge, can be used to obtain a causal model. While this can be a valuable aid in identification of confounders, there are still situations where this is insufficient to obtain an unbiased estimate,

specifically, when unmeasured confounders are involved (for example, the unobserved confounder U_4 in Supplemental Figure S1).

However, even then there are situations where it is still possible to obtain an unbiased estimate of effect of exposure on outcome. When an unconfounded mediator of treatment can be measured, one can estimate the effect in two steps: first the effect of the mediator on outcome and then the effect of treatment on the mediator. This construct is also known as the front door criterion.[13]

In oncology, a possibility would be to use recurrence-free survival (RFS) as a mediator for overall survival, assuming recurrence proceeds death due to cancer and is more likely to be driven by tumor and treatment characteristics and not influenced by confounders such as SES. While RFS is not routinely collected in the NCR, it is available for some tumors, and might be an interesting avenue to pursue.

Tables & Figures

Tables

variable	values	Active treatment vs observation: observation (n=1950)	active (n=2171)	p		p CMH	
age	66-69	708 (36.3)	890 (41.0)	0.0000	***	0.3351	
	70-74	710 (36.4)	895 (41.2)				
	75-80	532 (27.3)	386 (17.8)				
stage	I	1709 (87.6)	1620 (74.6)	0.0000	***	0.5450	
	II	241 (12.4)	551 (25.4)				
grade	1	57 (2.9)	51 (2.3)	0.2920		0.7411	
	2	1893 (97.1)	2120 (97.7)				
SES	low (1-3)	582 (29.8)	643 (29.6)	0.0615		0.8508	
	medium (4-7)	765 (39.2)	920 (42.4)				
	high (8-10)	603 (30.9)	607 (28.0)				
year of diagnosis	2005-2008	455 (23.3)	771 (35.5)	0.0000	***	0.7970	
	2009-2011	564 (28.9)	482 (22.2)				
	20012-2014	931 (47.7)	918 (42.3)				
PSA	1 – PSA < 10	913 (47.0)	776 (35.9)	0.0000	***	0000.0	***
	10 ≤ PSA < 20	754 (38.8	941 (43.6)				
	PSA ≥ 20	277 (14.2)	443 (20.5)				

Table 1: Variable distributions for active treatment and observation. 'p': calculated using Pearson's Chi2 test. 'p CMH': calculated using the Cochran-Mantel-Haenszel Chi2 test, stratifying by propensity score. Propensity scores were calculated using the variables *year of diagnosis*, *age*, *stage*, *grade*, and *SES*. *: $p \leq 0.05$, **: $p \leq 0.005$, ***: $p \leq 0.0005$.

Destination (child) \ Source (parent)	yod_cat	age	stage	SESq	PSAq	treatment	grade	surv_01y	surv_02y	surv_03y	surv_04y	surv_05y	surv_06y	surv_07y	surv_08y	surv_09y	surv_10y
yod_cat			X		X	X	X	X	X	X	X	X	X	X	X	X	X
age					X	X		X	X	X	X	X	X	X	X	X	X
stage						X		X	X	X	X	X	X	X	X	X	X
SESq						X		X	X	X	X	X	X	X	X	X	X
PSAq						X		X	X	X	X	X	X	X	X	X	X
treatment									X	X	X	X	X	X	X	X	X
grade			X			X		X	X	X	X	X	X	X	X	X	X
surv_01y									X	X	X	X	X	X	X	X	X
surv_02y		>>				>>		>>		X	X	X	X	X	X	X	X
surv_03y											X	X	X	X	X	X	X
surv_04y												X	X	X	X	X	X
surv_05y						>>							X	X	X	X	X
surv_06y														X	X	X	X
surv_07y															X	X	X
surv_08y																X	X
surv_09y		>															X
surv_10y	>					>>											

Table 2: Structure learning constraints for Adjuvant androgen deprivation vs none. The blacklist is made up from cells containing 'x': these edges are *forbidden*. Minimal and extended whitelists (i.e., *enforced* edges) are formed by cells marked with '>' and '>>' respectively.

						5-year survival			10-year survival		
		HR	95% CI	p	sign	No	Yes	Δ	No	Yes	Δ
1	Unadjusted	0.81	0.73-0.91	0.0004	***	0,88	0,90	0,02	0,65	0,71	0,05
2	Adjusted for year of diagnosis, age, SES, stage, and grade	0.87	0.77-0.98	0.0201	*	0,89	0,90	0,01	0,69	0,72	0,03
3	Adjusted for age, stage, grade, SES and propensity score	0.88	0.78-0.98	0.0267	*	0,88	0,90	0,01	0,67	0,70	0,03
4	Stratified by quintile of propensity score q1	0.94	0.72-1.23	0.6670		0,85	0,85	0,01	0,60	0,62	0,02
	q2	1.09	0.83-1.42	0.5305		0,91	0,91	-0,01	0,66	0,64	-0,02
	q3	0.70	0.53-0.94	0.0174	*	0,89	0,92	0,03	0,55	0,65	0,11
	q4	0.94	0.74-1.20	0.6217		0,90	0,91	0,01	0,72	0,74	0,01
	q5	0.66	0.52-0.85	0.0012	**	0,85	0,89	0,05	0,61	0,72	0,11
5	Adjusted for age and year of diagnosis	0.88	0.79-0.99	0.0340	*	0,89	0,90	0,01	0,67	0,70	0,03

Table 3: Hazard Ratios (HR) and confidence intervals (CI) for active treatment ("Yes") versus observation ("No", referent category) and probabilities for 5- and 10-year survival. Central values (statistical mode) used, where appropriate, were age: 66-69, SES: 2 - medium (4-7), year of diagnosis: 2012-2014, stage: I, grade: g2, propensity: q3. The highlighted row (nr. 5) was not part of the original analyses but identified as most appropriate by our causal model. *: $p \leq 0.05$, **: $p \leq 0.005$, ***: $p \leq 0.0005$.

Figures

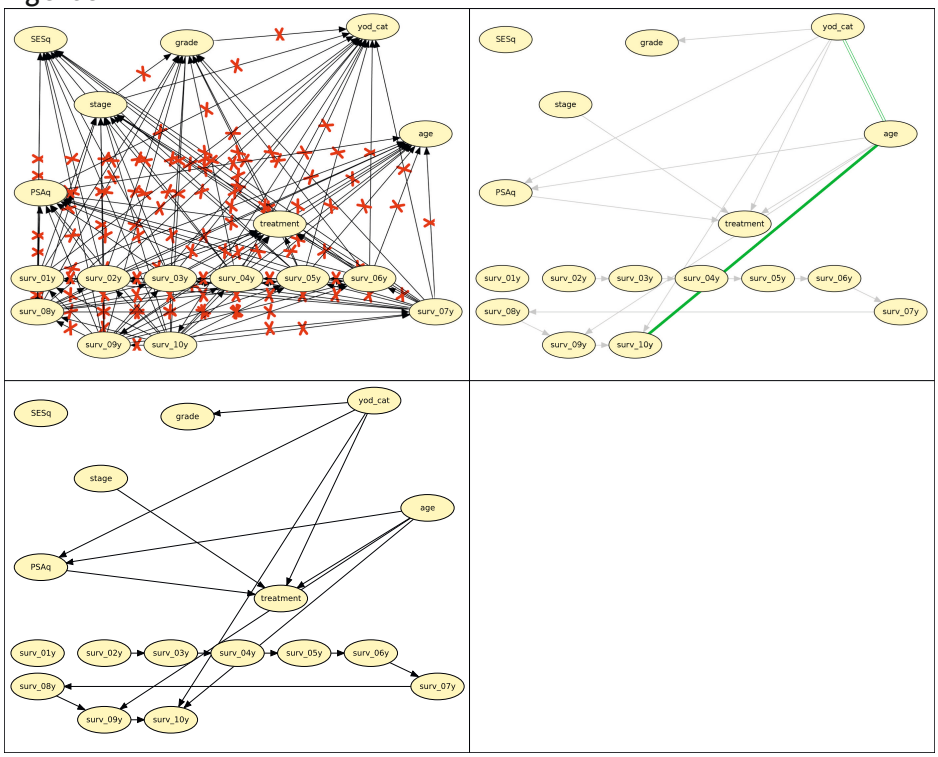


Figure 1: Structure learning process using a *blacklist and a minimal whitelist*. *Top left:* initial state, showing imposed constraints. *Top right:* ambiguous region identified by Hugin, which was resolved by selecting age → surv_10y. *Bottom left:* Final result. A larger image of this panel is available as Supplemental Figure S9.

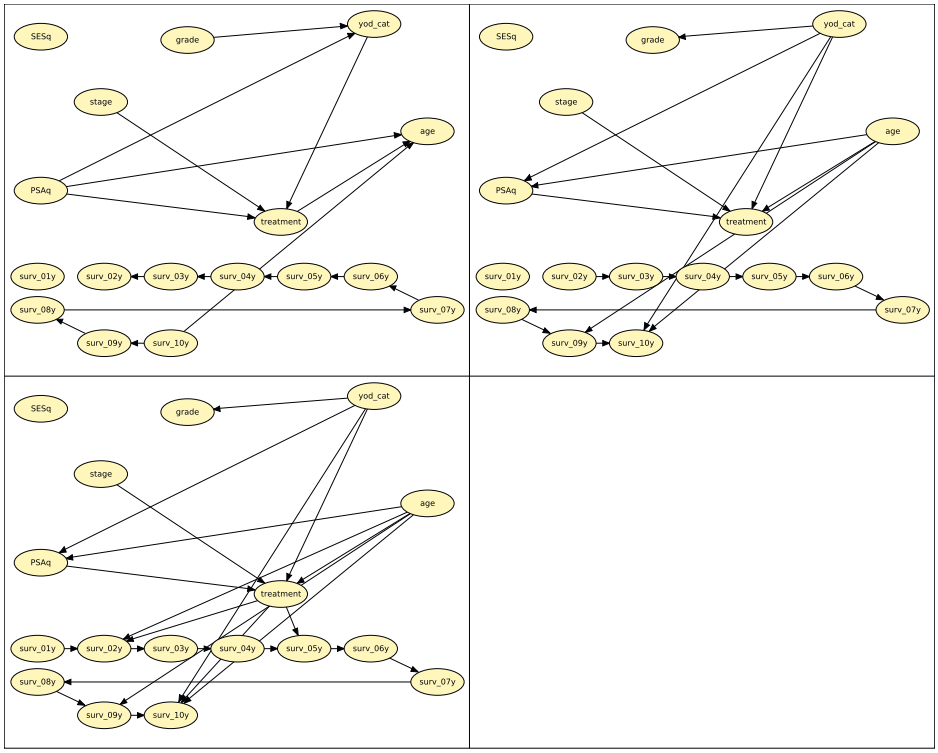


Figure 2: BN structures found using different structure learning constraints. Larger images are available in the supplemental materials. *Top left:* BN structure found by the NPC-algorithm when not providing any structure learning constraints (i.e., prior knowledge). In addition to the edges between the survival nodes, the edges grade → year of diagnosis (labelled "yod_cat"), PSA (labelled "PSAq") → age, PSA → year of diagnosis, treatment → age, and 10y-survival (labelled "surv_10y") → age can be considered non-causal. No edges were found between treatment and survival. *Top right:* BN structure found using a blacklist with a minimal whitelist. *Bottom left:* Final BN structure created using blacklist and extended whitelist. Year of diagnosis (labelled "yod_cat") and age act as confounders. Paths from PSA (labelled "PSAq") to survival go through either year of diagnosis or age.

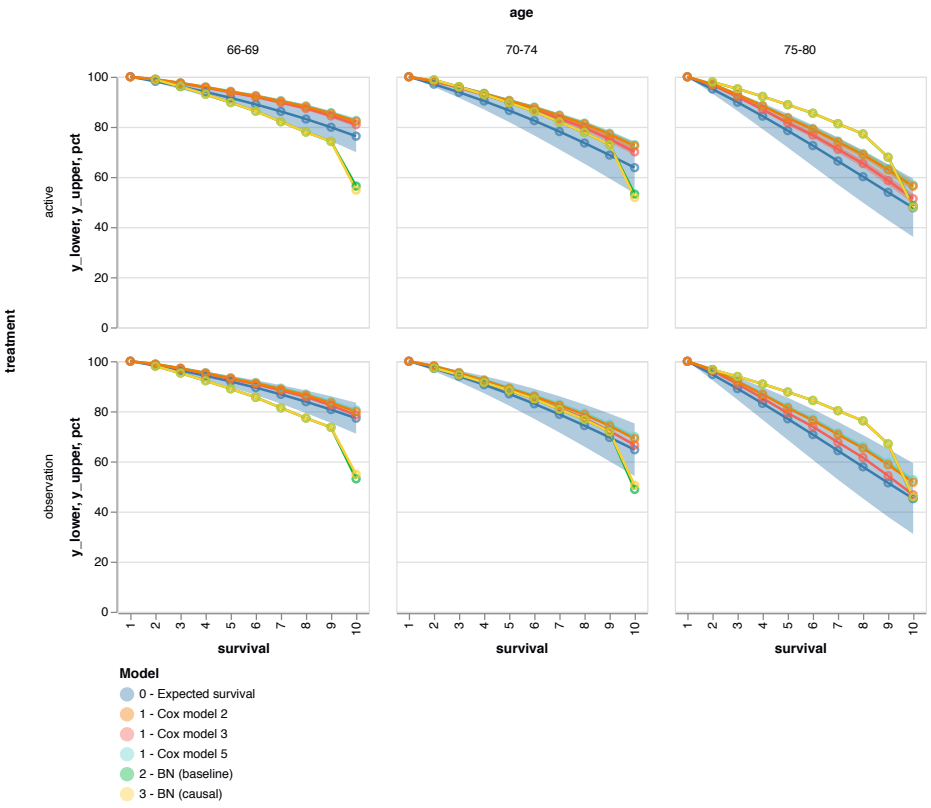


Figure 3: Predicted and expected 10-year survival for localized prostate cancer, stratified by treatment (rows), age (columns) and model (color). The Cox models in the top row (active treatment) have a very narrow CI which is (nearly) invisible in this plot.

References

1. van Maaren MC, Poortmans P, Siesling S: Breast-conserving therapy versus mastectomy. Oncoscience 3:304–305, 2016
2. Vernooij RWM, van Oort I, de Reijke TM, et al: Nationwide treatment patterns and survival of older patients with prostate cancer. J Geriatr Oncol 10:252–258, 2019
3. Dinmohamed AG, Cellamare M, Visser O, et al: The impact of the temporary suspension of national cancer screening programmes due to the COVID-19 epidemic on the diagnosis of breast and colorectal cancer in the Netherlands. J Hematol OncolJ Hematol Oncol 13:147, 2020
4. Keikes L, Koopman M, Lemmens VEPP, et al: Practice Variation in the Adjuvant Treatment of Colon Cancer in the Netherlands: A Population-based Study. Anticancer Res 40:4331–4341, 2020
5. Giordano SH, Kuo Y-F, Duan Z, et al: Limits of observational data in determining outcomes from cancer therapy. Cancer 112:2456–2466, 2008
6. Grimes DA, Schulz KF: Bias and causal associations in observational research. Lancet Lond Engl 359:248–252, 2002
7. Grimes DA, Schulz KF: False alarms and pseudo-epidemics: the limitations of observational epidemiology. Obstet Gynecol 120:920–927, 2012
8. Rubin DB, Thomas N: Matching using estimated propensity scores: relating theory to practice. Biometrics 52:249–264, 1996
9. Rubin DB: Estimating causal effects from large data sets using propensity scores. Ann Intern Med 127:757–763, 1997
10. VanderWeele TJ, Shpitser I: A new criterion for confounder selection. Biometrics 67:1406–1413, 2011
11. Häggström J: Data-driven confounder selection via Markov and Bayesian networks. Biometrics 74:389–398, 2018
12. Pearl J: Causal Diagrams for Empirical Research. Biometrika 82:669–688, 1995
13. Pearl J: Causality: Models, Reasoning and Inference (ed 2nd). New York, NY, USA, Cambridge University Press, 2009
14. Koller D, Friedman N: Probabilistic Graphical Models: Principles and Techniques. MIT Press, 2009
15. Wong Y-N, Mitra N, Hudes G, et al: Survival associated with treatment vs observation of localized prostate cancer in elderly men. JAMA 296:2683–2693, 2006
16. Mantel N: Chi-Square Tests with One Degree of Freedom; Extensions of the Mantel-Haenszel Procedure. J Am Stat Assoc 58:690–700, 1963
17. Mantel N, Haenszel W: Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. JNCI J Natl Cancer Inst 22:719–748, 1959
18. Agresti A: Categorical Data Analysis Second Edition. Wiley, New York, 2002
19. Andersen SK, Olesen KG, Jensen FV, et al: HUGIN - a Shell for Building Bayesian Belief Universes for Expert Systems6
20. Steck H: Constraint-based structural learning in Bayesian networks using finite data sets, in 2001
21. Dempster AP, Laird NM, Rubin DB: Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc Ser B Methodol 39:1–38, 1977
22. Sieswerda M: Thomas, a Python framework for working with Bayesian Networks. [Internet], 2021Available from: https://github.com/mellesies/thomas-core
23. Tsamardinos I, Brown LE, Aliferis CF: The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn 65:31–78, 2006
24. StatLine - Overlevingskansen; geslacht, leeftijd [Internet][cited 2022 May 17] Available from: https://opendata.cbs.nl/#/CBS/nl/dataset/70701ned/table?dl=68288
25. Schisterman EF, Cole SR, Platt RW: Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiol Camb Mass 20:488–495, 2009

26. Robinson LD, Jewell NP: Some Surprising Results about Covariate Adjustment in Logistic Regression Models. Int Stat Rev Rev Int Stat 59:227–240, 1991
27. Bill-Axelson A, Holmberg L, Ruutu M, et al: Radical prostatectomy versus watchful waiting in early prostate cancer. N Engl J Med 352:1977–1984, 2005
28. Hamdy FC, Donovan JL, Lane JA, et al: 10-Year Outcomes after Monitoring, Surgery, or Radiotherapy for Localized Prostate Cancer. N Engl J Med 375:1415–1424, 2016
29. Verma TS, Pearl J: On the Equivalence of Causal Models [Internet], 1990[cited 2021 Dec 21] Available from: https://arxiv.org/abs/1304.1108v1

Supplemental materials

Supplemental Tables

BN1	BN2	Structural Hamming Distance	Absolute difference/distance
No constraints	Blacklist only	7	14
	Minimal whitelist	9	15
	Extended whitelist	20	20
Blacklist only	Minimal whitelist	2	2
	Extended whitelist	7	7
Minimal whitelist	Extended whitelist	5	5

Table S1: Structural Hamming Distances and absolute differences between the BNs developed using a) no constraints, b) the minimal whitelist only and b) the extended whitelist.

observation	Active treatment vs observation		Difference (% point)
	observation	active	
P **(*survival* | *treatment*)**			
5-year survival	88.40%	89.51%	1.11
10-year survival	49.42%	53.66%	4.24
P **(*survival* | *do* (*treatment*))**			
5-year survival	88.45%	89.48%	1.02
10-year survival	51.13%	52.13%	1.00

Table S2: Probability of survival given treatment (i.e., *without* using causal reasoning) and given do(treatment) (i.e., *with* causal reasoning) as calculated by the BN. Probabilities that should match those calculated by Cox model 2 have been highlighted.

		This study				Giordano et al.			
		HR	CI	p	sign.	HR	CI	sign.	CIs overlap?
1	Unadjusted	0.81	0.73-0.91	0.0004	***	0.53	0.51-0.55	yes	no
2	Adjusted for year of diagnosis, age, SES, stage, and grade	0.87	0.77-0.98	0.0201	*	0.67	0.65-0.70	yes	no
3	Adjusted for age, stage, grade, SES, and propensity score	0.88	0.78-0.98	0.0267	*	0.68	0.65-0.70	yes	no
4	Stratified by quintile of propensity score q1	0.94	0.72-1.23	0.6670		0.71	0.66-0.76	yes	yes
	q2	1.09	0.83-1.42	0.5305		0.67	0.62-0.72	yes	no
	q3	0.70	0.53-0.94	0.0174	*	0.68	0.62-0.74	yes	yes
	q4	0.94	0.74-1.20	0.6217		0.64	0.56-0.72	yes	no
	q5	0.66	0.52-0.85	0.0012	**	0.57	0.48-0.67	yes	yes
5	Adjusted for age and year of diagnosis	0.88	0.79-0.99	0.0340	*				

Table S3: Comparison of Hazard Ratios (HR) and confidence intervals (CI) for active treatment ("Yes") versus observation ("No", referent category) between this study and the original analyses by Giordano et al. Central values (statistical mode) used for this study, where appropriate, were age: 66-69, SES: 2 - medium (4-7), year of diagnosis: 2012-2014, stage: I, grade: g2, propensity: q3. The highlighted row (nr. 5) was not part of the original analyses but identified as most appropriate by the causal model. *: $p \leq 0.05$, **: $p \leq 0.005$, ***: $p \leq 0.0005$.

Supplemental Figures

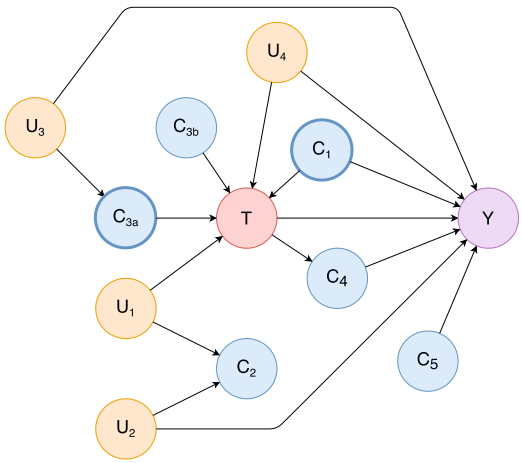


Figure S1: Different types of pathways between treatment (T) and outcome (Y). Blue nodes represent measured variables that could be considered potential confounders, orange nodes represent unmeasured variables. Correcting for nodes C_1 and C_{3a} is required to obtain an unbiased result; correction for unmeasured variables U_1, and U_2 is unnecessary. Adjusting for C_2 would open a backdoor path between T and Y, thus biasing the estimate. Correcting for C_{3a} and/or C_5 would be an unnecessary, bias-neutral adjustment. Correction for C_4 would yield unpredictable results, ranging from overestimating, to nullifying, to reversing any estimated effect. U_4 is the only unmeasured confounder that cannot be corrected for. Figure adapted from VanderWeele and Shpitser [10], Figure 1).

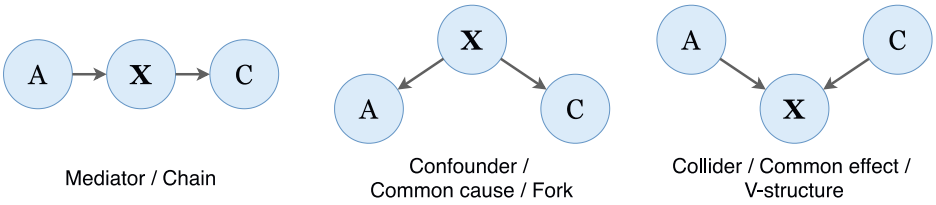


Figure S2: The three building blocks of causal models. The first two panels, mediator and common cause, model the same conditional *in*dependency: $A \perp\!\!\!\perp C \mid X$. The third panel models a conditional *de*pendency: $A \not\perp\!\!\!\perp C \mid X$

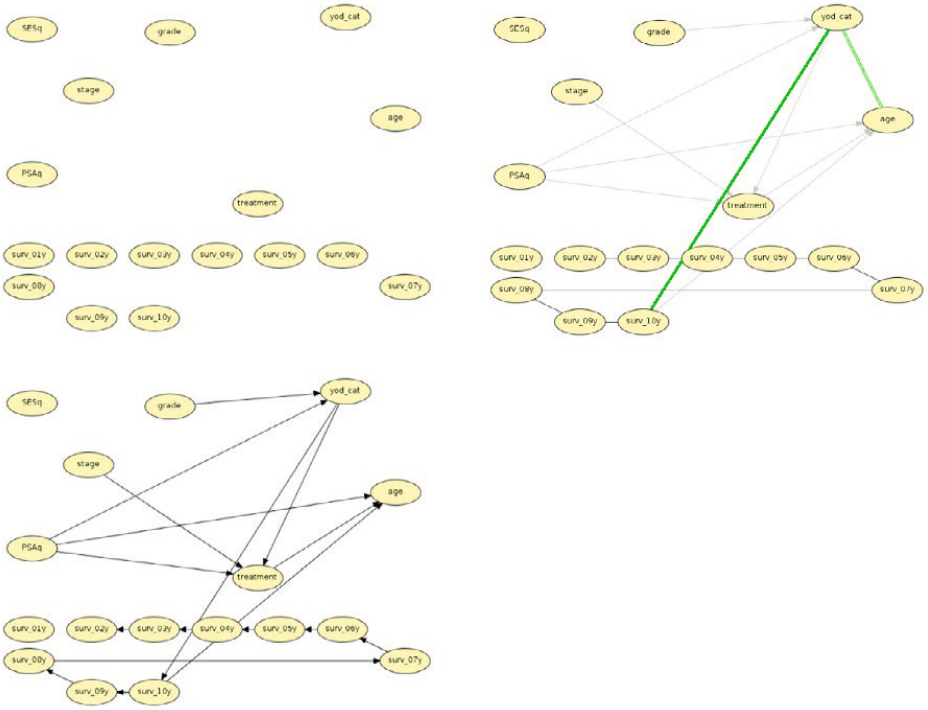


Figure S3: Structure learning process *without structure learning constraints*. *Top left:* Initial state. *Top right:* Edges identified by Hugin, including one ambiguous region which was resolved by selecting yod_cat - surv_10y. *Bottom left:* Final result after manual resolution of the ambiguous region. See Figure S7 for a larger image of this panel.

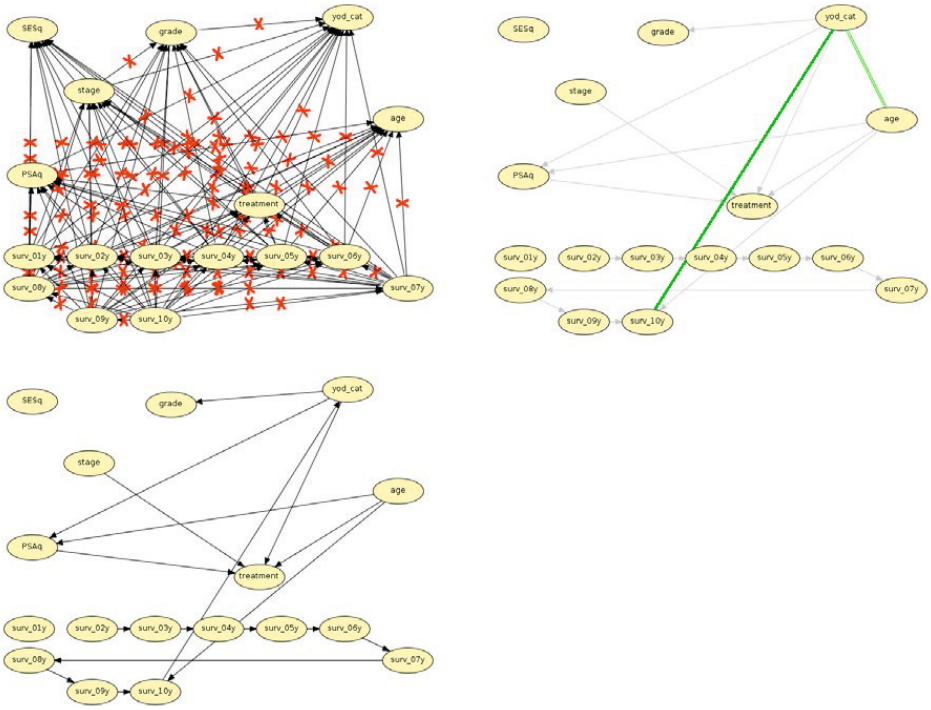


Figure S4: Structure learning process using a *blacklist only*. *Top left:* Initial state, showing imposed constraints. *Top right:* Edges identified by Hugin, including one ambiguous region (green edges) which was resolved by selecting yod_cat - surv_10y. *Bottom left:* Final result after manual resolution of the ambiguous region. Hugin directed the edge surv_10y → yod_cat against the blacklist. See Figure S8 for a larger image of this panel.

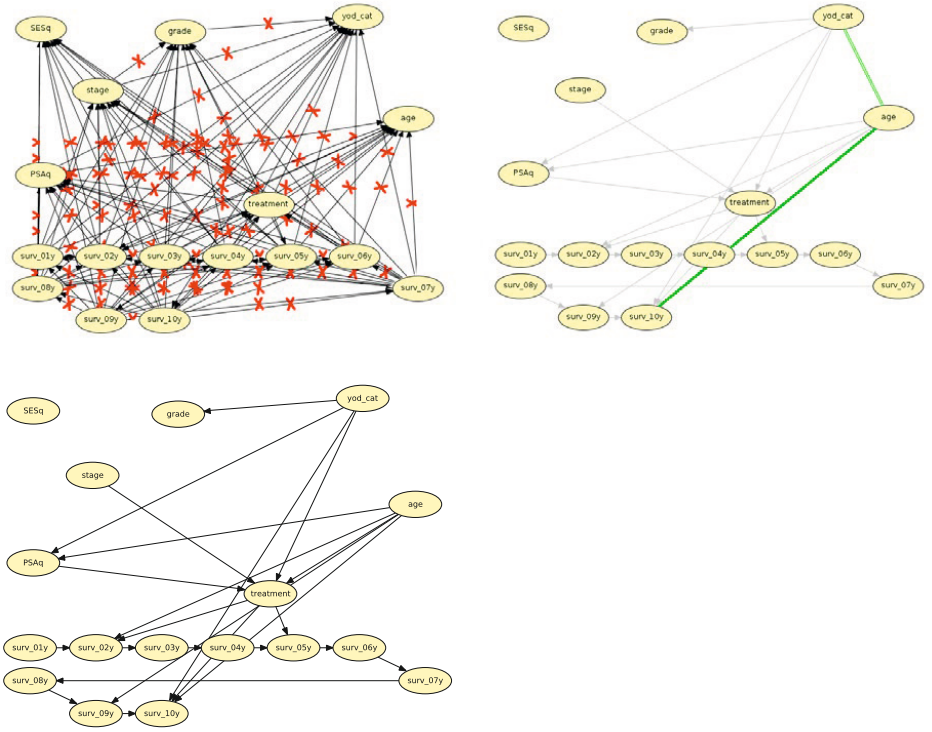


Figure S5: Structure learning process using a *blacklist and extended whitelist*. *Top left:* Initial state showing imposed constraints and enforced edges. *Top right:* Edges identified by Hugin, including one ambiguous region. *Bottom left:* Final result. See Figure S10 for a larger image of this panel.

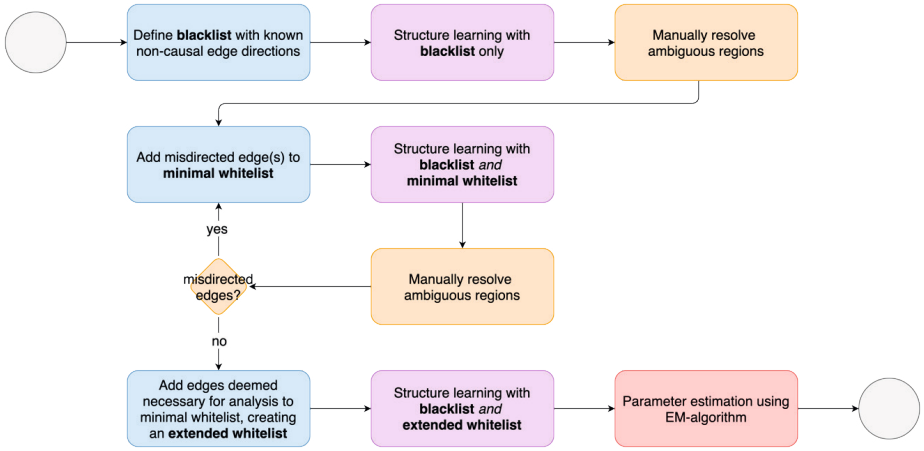


Figure S6: Overview of the process of Bayesian Network development using the NPC-algorithm, incorporating clinical knowledge.

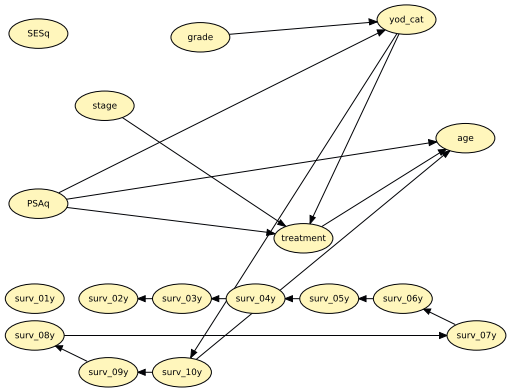


Figure S7: Final result of structure learning process *without structure learning constraints* (larger image). In addition to the edges between the survival nodes, the edges grade → year of diagnosis (labelled "yod_cat"), PSA (labelled "PSAq") → age, PSA → year of diagnosis, treatment → age, and 10y-survival (labelled "surv_10y") → age can be considered non-causal. No edges were found between treatment and survival.

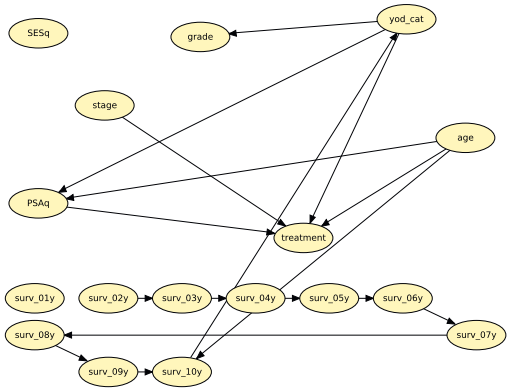


Figure S8: Final result of structure learning process using a *blacklist only* (larger image).

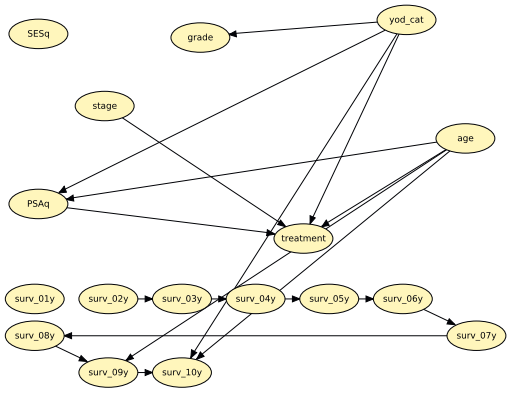


Figure S9: Final result of structure learning process using a *blacklist and minimal whitelist* (larger image).

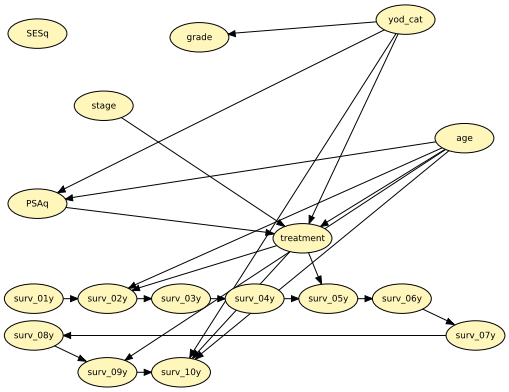


Figure S10: Final result of structure learning process using a *blacklist and extended whitelist* (larger image).

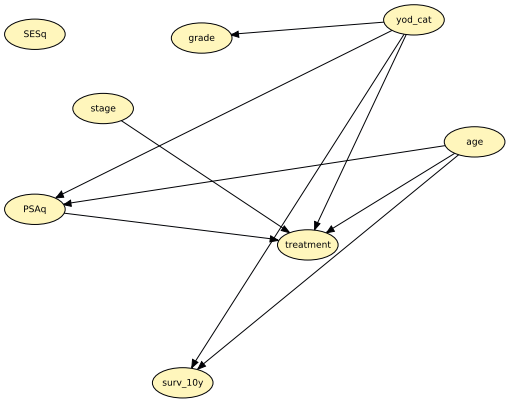


Figure S11: Network structure found by Hugin when using a simplified dataset (only 1 survival node) and blacklist only.

Supplemental to Introduction

Confounding

Here, we define a confounder as any variable with a causal effect on both treatment selection and outcome. The simplest example is a model with edges $X \rightarrow Y$, $X \leftarrow Z \rightarrow Y$ where X represents treatment, Y represents outcome and Z is the confounder (see also Supplemental Figure S12).

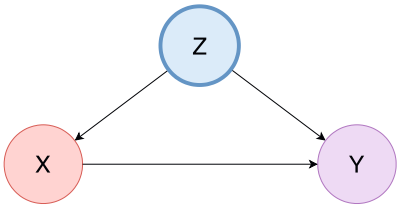


Figure S12: Simple causal model visualizing relationships between variables X, Y and Z. Here X represents treatment, Y represents outcome and Z is a (measured) confounder.

Graph equivalence

As stated in the introduction Bayesian Networks (BNs) are a type of probabilistic graphical model that use a directed acyclic graph (DAG) where nodes represent variables, and the directed edges signify (preferably causal) relationships.[13, 14] Each node is associated with a probability distribution that is conditional on its parents (i.e., the set of nodes that have a directed edge to that node. This implies a direct relationship between the edges in a network and the factorization of its distribution P: the network $A \rightarrow B$ implies $P(A)P(B \mid A)$.

Vice versa, depending on its associated conditional independencies, a single probability distribution P may be represented by multiple Directed Acyclic Graphs (DAGs). Intuitively this is easy to grasp considering Bayes' Theorem, $P(A)P(B \mid A) = P(A, B) = P(B)P(A \mid B)$, which allows an edge between A and B to be arbitrarily directed.

This concept can be extended to a more complex distribution consisting of multiple variables. For example, the 3 networks in Supplemental Figure S13 can be represented by the following equations:

1. $P(X)P(Y \mid X)P(Z \mid Y)$
2. $P(Y)P(X \mid Y)P(Z \mid Y)$
3. $P(Z)P(Y \mid X)P(Z \mid Y)$

Following Bayes' theorem $P(X)P(Y \mid X) = P(Y)P(X \mid Y)$, thus equations 1 and 2 are equal. Additionally, considering $P(Y)P(Z \mid Y) = P(Z)P(Y \mid Z)$, thus equations 2 and 3 are equal, which demonstrates overall equivalence.

The above is generalized by the concept of Markov equivalence. A Markov equivalence class is a set of DAGs that encode the same set of conditional independencies. Two DAGs are equivalent if and only if they have the same skeleton and the same colliders [29]. For example, the three networks $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \rightarrow Z$ and $X \leftarrow Y \leftarrow Z$ (Supplemental Figure S13) all capture the same conditional (in)dependencies: $X \perp\!\!\!\perp Z \mid Y$, $X \not\perp\!\!\!\perp Y$, and $Y \not\perp\!\!\!\perp Z$ and are Markov-equivalent.

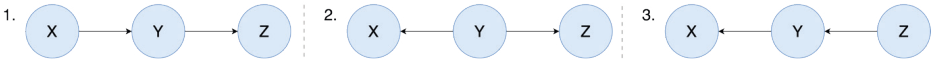


Figure S13: Three networks that capture the same probability distribution with its associated conditional independencies and can be considered Markov-equivalent.

Supplemental to Methods

Breakdown of applying in-/exclusion criteria

step	description	n	delta
0	Male prostate cancer patients, diagnosed between [2005, 2014] with 66 ≤ age ≤ 80, cT ≤ 2C and cM0	31577	
1	Dropping primary HTx, CTx, CRTx, Targeted Tx, etc.	26005	5572
2	Keeping records marked as “active treatment” or “observation” only	25090	915
3	Dropping records where grade is NA	18067	7023
4	Dropping cNX	7812	10255
5	Dropping cN in [“1”, “1M”]	7676	136
6	Dropping grade > 2	4192	3484
7	Dropping patients with survival < 12m	4121	71

Table S4: Results of applying in- and exclusion criteria on the dataset.

Determining active treatment

For determining primary treatment, two variables were considered: treatment sequence number and date. In about 10% date of first treatment was missing. Since registration takes place around 9 months after diagnosis and the date was within 6 months in ~96% of the records where it was available, events with missing dates were considered to have taken place within 6 months of diagnosis.

Primary radiotherapy (both external and internal) and prostatectomy were classified as active treatment. If first treatment consisted of lymph node dissection, the case was included as active treatment only if subsequent treatment could be considered active treatment *and* took place within 6 months of diagnosis. Treatment specifically marked as “no treatment” was classified as observation, as were 670 cases where (any) treatment started later than 180 days after diagnosis. Patients that underwent primary hormonal

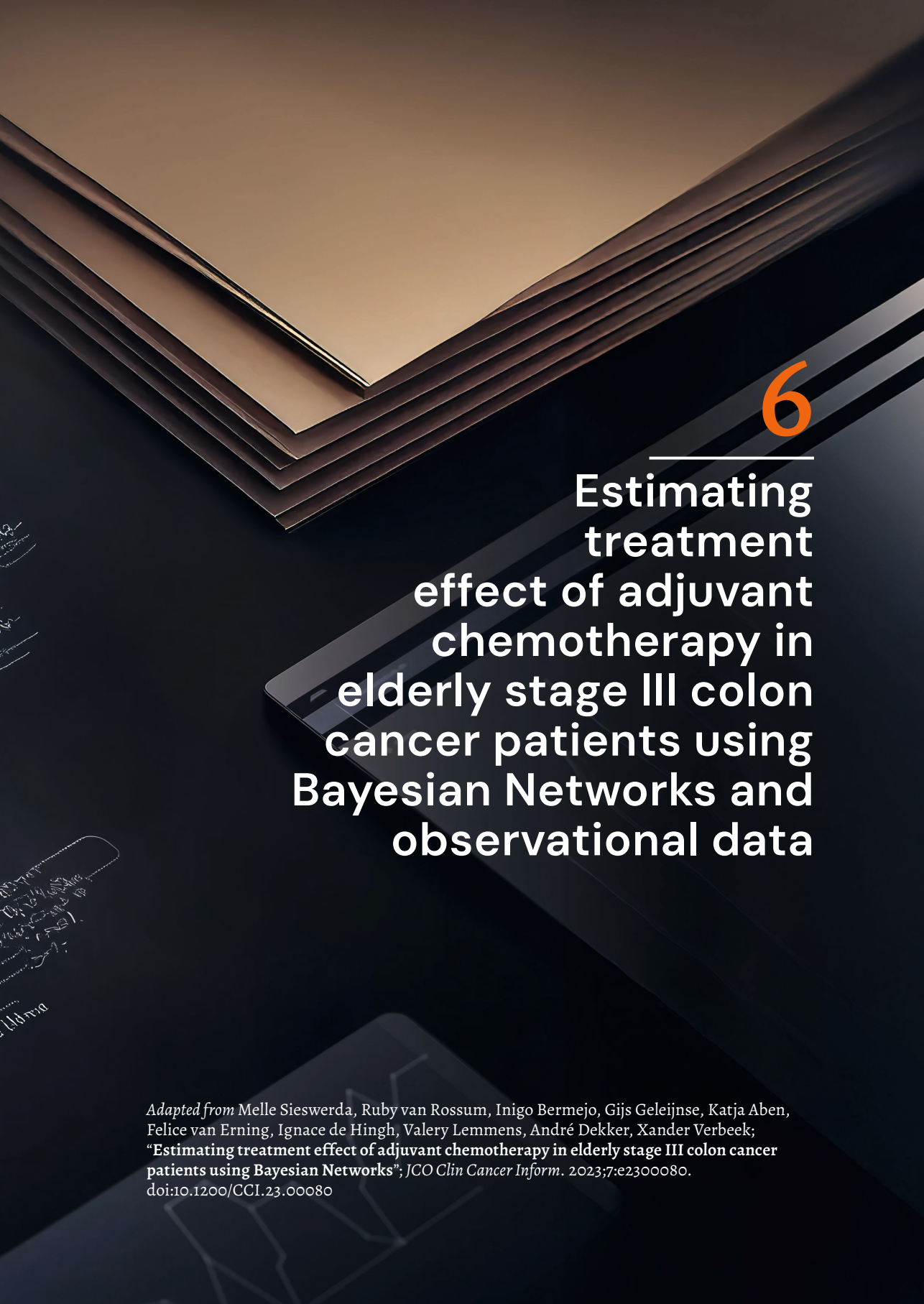
therapy, chemotherapy, targeted therapy, or radiotherapy on metastases were excluded, as were patients that died within 1 year of diagnosis.

Bayesian network development

In general, selection of clinical knowledge may lead to bias. For example, PSA is often (implicitly) considered to be the result of disease (and thus used as a marker of disease progression). This could lead to the decision to preclude edges from PSA to stage (or grade), but would rule out the possibility in the model that PSA *contributes* to disease progression.

In certain situations, it may not even be possible to identify the counterfactual relationship, because two variables are part of a feedback system (does pain cause/worsen fatigue or does fatigue worsen pain?) or children to a common, but unmeasured, parent (PSA and stage being, once again, an example).

In this case, however, we feel definitions of the precluded edges were conservative and unambiguous. Still, we've listed the imposed constraints and would argue that making these assumptions explicit will foster discussion and, hopefully, lead to better understanding of disease processes & treatment effects.



6

Estimating treatment effect of adjuvant chemotherapy in elderly stage III colon cancer patients using Bayesian Networks and observational data

Adapted from Melle Sieswerda, Ruby van Rossum, Inigo Bermejo, Gijs Geleijnse, Katja Aben, Felice van Erning, Ignace de Hingh, Valery Lemmens, André Dekker, Xander Verbeek; "**Estimating treatment effect of adjuvant chemotherapy in elderly stage III colon cancer patients using Bayesian Networks**"; *JCO Clin Cancer Inform*. 2023;7:e2300080. doi:10.1200/CCI.23.00080

Abstract

PURPOSE: While adjuvant therapy with capecitabine and oxaliplatin (CAPOX) has been proven effective in stage III colon cancer, capecitabine monotherapy (CapMono) might be equally effective in elderly patients. Unfortunately, the elderly are underrepresented in clinical trials and patients included may not be representative of the routine care population. Observational data might alleviate this problem but is sensitive to biases such as confounding by indication. Here, we build causal models using Bayesian Networks (BNs), identify confounders, and estimate the effect of adjuvant chemotherapy using survival analyses.

METHODS: Patients aged ≥ 70y were selected from the Netherlands Cancer Registry (N=982). We developed several BNs using constraint-based, score-based, and hybrid algorithms while precluding non-causal relations. Additionally, we created models using a limited set of recurrence and survival nodes. Potential confounders were identified through the resulting graphs. Several Cox-models were fitted correcting for confounders, as well as for propensity scores.

RESULTS: When comparing adjuvant treatment to surgery only, pN, physical status, and age were identified as potential confounders. Adjuvant treatment was significantly associated with survival in all Cox models, with hazard ratios between 0.39 and 0.45; Confidence Intervals overlapped. BNs investigating CAPOX vs CapMono did not find any association between treatment choice and survival, thus no confounders. Analyses using Cox-models did not identify significant association either.

DISCUSSION: We were able to successfully leverage BN structure learning algorithms in conjunction with clinical knowledge to create causal models. While confounders differed depending on the algorithm and included nodes, results were not contradictory. We found a strong effect of adjuvant therapy on survival in our cohort. Additional oxaliplatin did not have a marked effect and should be avoided in elderly patients.

Introduction

With over 8.000 new cases annually, which constitutes ~7% of the total cancer incidence, colon cancer is the fourth most common type of cancer in the Netherlands.[1] Given that treatment is always a balance between cost (both financial and physical) and benefit, it is important to continuously assess and improve the effect of (adjuvant) treatment.

For stage III colon cancer, Dutch guidelines recommend treatment with surgery and adjuvant chemotherapy with either capecitabine-oxaliplatin (CAPOX) or folinic acid-fluorouracil-oxaliplatin (FOLFOX). However, in elderly patients, the benefit of oxaliplatin is still a point of contention.[2, 3]

While 59% of the colon cancer patients are over 70 years old, elderly patients are frequently excluded from clinical trials based on age alone. Even if age does not rule out participation, elderly patients included may still not be representative for daily clinical practice as they must be fit enough to satisfy other inclusion criteria.

To bridge this gap, observational data plays a vital role. Unfortunately, when it comes to estimating treatment effect, observational data are sensitive to bias, specifically selection biases like confounding by indication. For example, if predominantly relatively healthy patients receive treatment, this will skew results in favor of treatment.

There are different ways to account for confounding by indication, for example through including confounders as covariates (in regression analysis), stratification, or propensity score correction.[4–9] In any case, correction requires a decision on which variables to include as confounders. Several criteria exist to help in this decision, for example the pre-treatment, common-cause, backdoor-path, and disjunctive cause criterion.[10–12] The pre-treatment criterion would select *all* pre-treatment variables as confounders. The common-cause criterion would, as the name suggests, only correct for those variables thought to be common causes of exposure and outcome. The backdoor-path criterion resembles the common-cause criterion but takes chains of influence into account and can thus correct for indirect confounders. The disjunctive cause criterion would correct for those variables thought to be either a cause of exposure or outcome.

When there is uncertainty about the presence of unmeasured confounding, it can be reasoned that, generally, the disjunctive cause or backdoor-path criteria yield the most unbiased results. For example, correcting for variable C_2 in Figure 1, which would be corrected for when applying the pre-treatment criterion, would unnecessarily bias the estimated effect of $T \rightarrow Y$. Application of these criteria, however, requires a causal model. Interestingly, structure learning algorithms for Bayesian Networks (BNs) can be used – with input from domain experts – to discover causal models from data, as recently shown.[13]

Briefly, BNs are a type of probabilistic graphical model whose structure is determined by a directed acyclic graph where nodes represent variables and directed edges signify (preferably causal) relationships.[11, 14] Each node is associated with a probability distribution conditional on its parents (i.e., nodes with a directed edge toward that node).

Structure learning algorithms for BNs can be classified into three categories. Constraint-based algorithms (e.g., NPC: Necessary Path Condition[15]) leverage the presence of conditional (in)dependencies to determine a model that best fits the data, score-based algorithms (e.g., SM: Silander-Myllymaki[16]) optimize a heuristic that describes goodness of fit, and hybrid algorithms (e.g., MMPC: Max-Min Parents-and-Children[17]) combine constraint- and score-based strategies.

Generally, these algorithms can incorporate prior (clinical) knowledge in the form of mandatory and prohibited (directed) edges. And while causal relations (mandatory edges) between variables may be contested, counterfactuals (prohibited edges) can often be supplied with reasonable certainty and consensus. For example, tumor characteristics, treatment, or outcome will never *cause* age even if the reverse relation (e.g., age → tumor characteristics) is contended.

When using structure learning in conjunction with these expert defined counterfactuals (a blacklist consisting of prohibited edge directions) it is possible to obtain a causal graph. This model can then be used for confounder identification by applying one of the aforementioned criteria and subsequently mitigation.

Previously, we analyzed the effect of adjuvant chemotherapy with capecitabine and oxaliplatin (CAPOX) or capacitabin monotherapy (CapMono) in elderly stage III colon cancer patients using data from the Netherlands Cancer Registry (NCR) which covers the entire Dutch population.[2] In this analysis, we adjusted for confounding by indication using the pre-treatment criterion.

In this paper, we again investigate treatment effect of adjuvant chemotherapy and the added benefit of oxaliplatin (CAPOX) over CapMono. To this end, we first build causal models using Bayesian Networks in conjunction with different structure learning algorithms and a varying number of recurrence/survival nodes. Finally, we explore the effect of correcting for different sets of confounders, as identified by these causal models, and compare this with propensity score correction.

Methods

Data

The original dataset from the study by Van Erning et al.[2] was used and vital status was updated up until January 2021. Median follow-up for overall survival (OS) was 58 months (after surgery) for all patients and 113 months for patients alive at last follow-up. Median follow-up for recurrence-free survival was 19 months (after surgery) and 29 months for patients alive at last follow-up.

Briefly, a cohort of patients, diagnosed in the southeastern part of the Netherlands between 2005-2012 with pathological stage III ($pT_{1-4}N_{1-2}M_0$) colon cancer and age ≥ 70 years was selected from the NCR. The variables sex, age, ASA-classification (a system indicating physical performance status, developed by the American Society of Anesthesiologists), pT, pN, tumor subsite (coded according to ICD-O-3), and differentiation grade were included. Additionally, in 2013 and 2014, details regarding adjuvant therapy, number of comorbidities and development of (local) recurrence were acquired from medical records and added to the NCR. Year of diagnosis was available but was dropped after a quick analysis of its association with recurrence (see Supplemental Materials – "Variable Selection" and Supplemental Materials – Figure S1).

As before, patients that died within 90 days after surgery (n=125) were excluded, since these deaths were likely due to surgical complications and patients were unable to undergo adjuvant chemotherapy. Patients receiving chemotherapy other than CAPOX or CapMono (e.g. FOLFOX) were excluded. This left a final dataset consisting of n=982 records (Figure 2, Table 1, Supplemental Materials – Tables S1 and S2). For analysis of CAPOX compared to CapMono, a subset consisting of patients that received adjuvant therapy was created (n=352).

For assessing association with treatment and for use in the BN, age was discretized into 3 categories: 70-74, 75-79, and ≥ 80 (see also Supplemental Materials – "Discretization of age" and Supplemental Materials – Figure S2). The anatomical subsite was relabeled as proximal (caecum – splenic flexure of colon), distal (descending – sigmoid colon), or unknown/unspecified.

OS, counted from date of surgery, was discretized into five Boolean variables representing 1- to 5-year survival, introducing NAs (missing values) for patients who had a follow-up < 5 years and were alive at the time of follow-up. Absence of recurrence was discretized similarly. However, due to the limited follow-up, variables representing absence of recurrence ≥ 3y were skewed toward recurrence or NA. Consequently, these three variables were dropped (see Supplemental Materials – "Variable Selection").

Statistical Analyses

Bayesian Network Development

Using different combinations of outcome variables full and simplified BNs were developed. The full networks used two nodes/variables to represent recurrence at 1 and 2 years, and five nodes for overall survival at 1, 2, 3, 4, and 5 years. The simplified networks used the last node of each.

The rationale for also evaluating a simplified network is the following. Having multiple outcome nodes at sequential time points might limit the BNs ability to find other associations. For example, death at 1 year after diagnosis perfectly predicts death for all subsequent nodes/timepoints. Therefore, when testing for association between any other variable and a survival node at a later time point, one is essentially testing for association with survival in a specific interval. It can be reasoned that this reduces the (statistical) power to find associations with later timepoints. E.g., if variable 1 is strongly associated with 1 year survival than any association of variable 2 with 2-year survival will be harder to detect. In the simplified networks only the final timepoint is included to obtain the maximal number of associations with the highest statistical power as the treatment effect on survival is expected to be the highest at that timepoint.

Additionally, we evaluated the effect of using different algorithms than the constraint-based NPC-algorithm, as available in Hugin.[18] Specifically, we applied MMHC (hybrid) and SM (score-based) as implemented in the R package bnstruct.[16, 17, 19, 20] Either adjuvant treatment ("Yes", "No") or adjuvant treatment regimen ("CAPOX", "CapMono") was included as appropriate. Obtained BNs were visualized using the R package qgraph.[21]

In all cases, structure learning was performed with level of significance set to 0.05.[15] In Hugin the process was supported using a manually defined blacklist of prohibited edge directions (Figure 3). Bnstruct uses a different, less granular approach and accepts a list of layers, where edges from lower to higher layers are forbidden.

The NPC-algorithm may yield ambiguous regions, consisting of "a set of inter-dependent uncertain links" where the absence of one link depends on the presence of another. When this happens, Hugin prompts for input. For resolution we first prioritized associations with outcome nodes (recurrence and survival) and subsequently associations with treatment.

In some situations, Hugin ignores the supplied blacklist after manual resolution and "misdirects" the selected edge. For example, Hugin created an edge from the node representing 4-year survival (surv_04y) toward the ASA node. To overcome this issue, misdirected edges were added to a (minimal) whitelist of enforced edges and structure learning was repeated.

The resulting graphs were inspected to identify confounders.

Propensity Score Analysis

Propensity scores were calculated as follows. First, the Chi^2-test was used to assess the association between pre-treatment variables and adjuvant treatment. Next, variables with a significant association with treatment were used as covariates in the propensity score calculation. For adjuvant treatment vs surgery only these were sex, age, comorbidities, ASA, and pN. For CAPOX vs CapMono these were age and comorbidities only.

Subsequently, the propensity score was discretized using quintiles. Finally, the Cochran-Mantel-Haenszel (CMH) Chi^2-test, with the bins as strata, was used to evaluate the effect of propensity score correction.[8, 9, 22]

Cox Proportional Hazards Models

Several Cox models were developed, each investigating the effect of (regimen of) adjuvant treatment on survival while correcting for a different set of potential confounders, driven by the results from structure learning. Covariates included are listed in Figure 4. Hazard ratio's (HRs) for treatment were extracted, together with their confidence intervals (CIs).

Results

Adjuvant Treatment vs Surgery Only

Variable Distribution and Propensity Score Correction

Chi^2 analysis showed variables age, sex, pN, ASA (physical performance status), and comorbidities were significantly associated with receiving any form of adjuvant treatment (Table 1, Supplemental Materials – Table S1). Stratification using the discretized propensity score as strata in the CMH Chi^2-test removed all associations between these conceivable confounders and treatment.

Bayesian network development & confounder identification

The full BNs are shown in the left column of Figure 5. In the BN found by the NPC-algorithm two confounders were identified through inspection of the graph: pN and ASA. The BNs found by MMHC- and SM-algorithms did not identify any confounders and generally had fewer edges.

In the simplified networks, which focused on a limited set of recurrence and survival nodes (Figure 5, right column), the NPC-algorithm again identified pN as confounder but ASA, which was associated with surv_04y in the full network, was no longer associated with survival. The SM-algorithm did not find an association between pN and treatment

(node adj_therapy) but *did* find an association between age and surv_05y, identifying age as a potential confounder.

In summary, depending on the algorithm selected and nodes included, treatment was either unconfounded or confounded by pN, pN and ASA, or age.

Cox Proportional Hazards Models

Adjuvant treatment was significantly associated with survival in all models (Figure 4, top panel; Supplemental Materials - Table S3). Estimated HRs were between 0.39 and 0.45, depending on covariates included. The models had overlapping CIs and yielded comparable results.

CAPOX vs CapMono

Variable Distribution and Propensity Score Correction

Chi^2 analysis showed variables age and comorbidities to be significantly associated with the choice between CAPOX or CapMono (Table 1, Supplemental Materials – Table S2). Here too, stratification using the binned propensity score as strata in the CMH Chi^2-test removed all associations between these suspected confounders and treatment.

Bayesian network development & confounder identification

From the full networks, none of the algorithms identified an association between adjuvant treatment regimen and survival (Figure 6). From the pre-treatment variables only grade was associated with survival (using the NPC-algorithm) and no confounders were found.

In the simplified networks, again, none of the algorithms identified an association between adjuvant treatment regimen and survival. All networks suggested that age was a driving factor for choice of regimen. The SM-algorithm implied that pN was an independent predictor of survival.

Summarized, regardless of the algorithm selected and nodes included, choice of treatment regimen was unconfounded, but also unassociated with survival.

Cox Proportional Hazards Models

None of the Cox models found significant association between adjuvant treatment regimen and survival with HR point estimates around 0.93 and CIs equal to or wider than 0.68 - 1.28 (Figure 4, bottom panel; Supplemental Materials – Table S4). Confidence intervals between the models were largely overlapping.

Discussion

Introduction

We were able to successfully leverage BN structure learning algorithms in conjunction with (basic) clinical knowledge to create causal models and subsequently identify potential confounders. Different structure learning algorithms identified different potential confounders, but none were contradicting. In other words, confounders identified in one model were not mediators or colliders in others (see Supplemental Materials - Figure S3) which may introduce bias if corrected for. From a theoretical point of view, score-based algorithms run a higher risk of incorrectly classifying a variable as confounder in the presence of *unmeasured* confounding (see Supplemental Materials - Figure S4). In this case this mechanism does not seem to play a significant role, since all HRs were very close together.

Adjuvant Treatment vs Surgery Only

When comparing adjuvant treatment with surgery only, the set of identified confounders differed, depending on both the specific algorithm used and the nodes included in the network. Potential confounders identified were pN, ASA, and age, although 3 (out of 6) networks suggested treatment effect was unconfounded. All Cox-models found significant benefit of adjuvant treatment over surgery only (HRs ranging between 0.36 – 0.46); correcting for different sets of potential confounders had no marked effect and neither did correction using propensity score.

CAPOX vs CapMono

In the comparison between CAPOX and CapMono, no confounders were identified, regardless of algorithm or nodes included: none of the BNs found an association between choice of treatment regimen and survival. The unadjusted Cox-model yielded the same conclusion and did not find any effect of adding oxaliplatin. Correcting for propensity score did not make a difference.

Associations found between pre-treatment variables

Looking at the BNs in general, several interesting associations between pre-treatment variables are found in all networks, regardless of algorithm or node selection. Unsurprisingly, age appears associated with the choice of adjuvant treatment. Not only when considering whether to start either regimen, but also when making a choice for a specific regimen. Similarly, sex was associated with tumor location and grade with pN, both associations that have been reported before.[23, 24] The association found between number of comorbidities and ASA-score also makes sense considering the use of the ASA-score (physical performance status) to approximate comorbidity in registries.[25] On the other hand, some expected associations were not, or only infrequently, found: only one network connected pT and survival. However, this might be explained by the fact

that we selected stage III and the (vast) majority of patients had a pT3 tumor. Overall, we considered the obtained networks clinically plausible.

Comparison with existing literature

Previously, Wilkinson et al. investigated the effect of adjuvant 5-fluorouracil with leucovorin (5-FU/LV) compared to surgery only.[26] The HR found here is smaller than the effect they reported, which was 0.62 in stage II – III and 0.65 in stage III only. It should be noted that the age distribution between our study (100% over 70 years old) and Wilkinson et al (~36-51% < 60y depending on treatment group) is vastly different. Our results seem to suggest that, in our elderly cohort, chemotherapy improves survival relatively more.

The effect of adding oxaliplatin to a fluoropyrimidine (e.g. capecitabine or fluorouracil) in stage II – III colon cancer has been extensively studied.[3, 27–32] Three trials in particular have contributed to our understanding: the NO16968 (XELOX in Adjuvant Colon Cancer Treatment) study, the Multicenter International Study of Oxaliplatin/5-Fluorouracil/ Leucovorin in the Adjuvant Treatment of Colon Cancer (MOSAIC) trial, and the C-07 trial of the National Surgical Adjuvant Breast and Bowel Project (NSABP).

NO16968 and MOSAIC found a small (4-6 percentage-point) but significant benefit of adding oxaliplatin to the treatment regimen. In MOSAIC, treatment effect was stronger in patients with stage III compared to stage II; NO16968 only targeted stage III patients.[29, 33] Both trials included a relatively young population. NSABP, which also investigated a younger population compared to our analysis, did not find an overall benefit of adding oxaliplatin but did report a significant effect in an unplanned subset analysis of patients ≤ 70 years old.[32] In a subgroup analysis, the MOSAIC study drew the complimentary conclusion that there is no additional benefit of oxaliplatin in elderly patients. This is in line with the results obtained here.

Conclusion

In conclusion, we have shown that structure learning elucidates underlying relationships in data, helping select which variables should be corrected for. Varying included variables and algorithms yielded (slightly) different, complementary results and identified different sets of potential confounders. Hazard ratios were similar regardless of the chosen set.

We found a strong association between adjuvant treatment with capecitabine and survival in stage III colon cancer in our cohort of patients of ≥ 70 years old. No additional benefit of adding oxaliplatin was found. As such, addition of oxaliplatin may be considered in younger patients with more advanced stage but should be avoided in elderly patients.

Tables & Figures

Figure 1 – Pathways between treatment and outcome

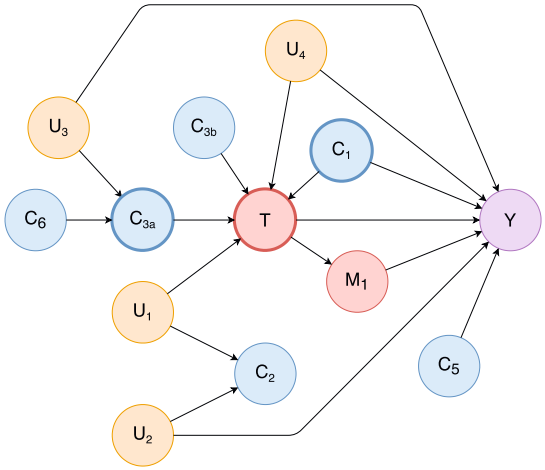


Figure 1: Different types of pathways between treatment (T) and outcome (Y). Blue nodes (C∗) represent measured variables that could be considered potential confounders, orange nodes (U∗) represent unmeasured variables. Correcting for nodes C_1 and C_{3a} is required to obtain an unbiased result. Correction for C_{3a} and C_6 simultaneously would open the backdoor path through U_3 thus should be avoided. Correction for unmeasured variables U_1, and U_2 is unnecessary. Adjusting for C_2 would open a backdoor path between T and Y, thus biasing the estimate. Correcting for C_{3b} and/or C_5 would be an unnecessary, bias-neutral adjustment. Correction for M_1 would yield unpredictable results, ranging from overestimating, to nullifying, to reversing any estimated effect. U_4 is the only unmeasured confounder that cannot be corrected for. Figure adapted from VanderWeele and Shpitser [12], Figure 1).

Figure 2 – CONSORT diagram

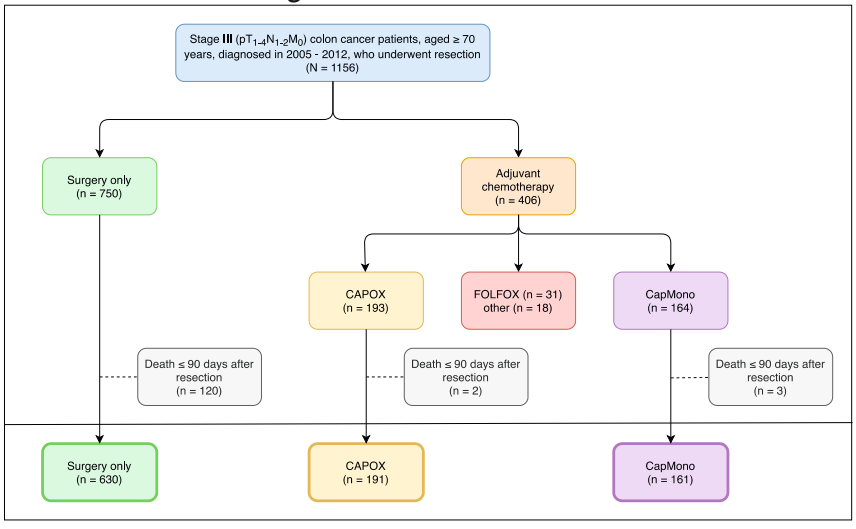


Figure 2: CONSORT diagram - Overview of patients included in the study.

Table 1 – Variable distribution

		Adjuvant therapy type			Adjuvant therapy vs Surgery Only			CAPOX vs CapMono		
variable	values	None	CapMono	CAPOX	p	sign.	p (CMH)	p	sign.	p (CMH)
sex	male	284 (45.1)	75 (46.6)	108 (56.5)	0.044	*	0.540	0.079		0.094
	female	346 (54.9)	86 (53.4)	83 (43.5)						
age	70 - 74	108 (17.1)	52 (32.3)	140 (73.3)	0.000	***	0.106	0.000	***	0.129
	75 - 79	197 (31.3)	84 (52.2)	47 (24.6)						
	80+	325 (51.6)	25 (15.5)	4 (2.1)						
co-mor-bidities	none	97 (15.4)	33 (20.5)	65 (34.0)	0.000	***	0.955	0.043	*	0.291
	1	135 (21.4)	52 (32.3)	48 (25.1)						
	2+	377 (59.8)	72 (44.7)	74 (38.7)						
	unknown	21 (3.3)	4 (2.5)	4 (2.1)						
ASA	1	16 (2.5)	12 (7.5)	25 (13.1)	0.000	***	0.513	0.141		0.220
	2	241 (38.3)	83 (51.6)	106 (55.5)						
	3	220 (34.9)	27 (16.8)	20 (10.5)						
	4	4 (0.6)	0 (0.0)	1 (0.5)						
	unknown	149 (23.7)	39 (24.2)	39 (20.4)						
pT	T1	10 (1.6)	6 (3.7)	4 (2.1)	0.259		0.157	0.286		0.400
	T2	51 (8.1)	18 (11.2)	18 (9.4)						
	T3	452 (71.7)	107 (66.5)	144 (75.4)						
	T4	117 (18.6)	30 (18.6)	25 (13.1)						
pN	N1	478 (75.9)	107 (66.5)	124 (64.9)	0.001	**	0.720	0.849		0.351
	N2	152 (24.1)	54 (33.5)	67 (35.1)						

Table 1: Variable distribution and association between variables and treatment. For Adjuvant Therapy vs Surgery only, propensity scores were calculated using *sex, age, comorbidities, ASA,* and *pN*. For CAPOX vs CapMono, *age* and *comorbidities* were used. 'p': calculated using Pearson Chi²-test. 'p*' calculated using the Cochran-Mantel-Haenszel Chi²-test after stratification by propensity score quintile. *: $p \leq 0.05$, **: $p \leq 0.005$, ***: $p \leq 0.0005$.

location	proximal	393 (62.4)	99 (61.5)	98 (51.3)	0.142	0.991	0.159	0.428
	distal	230 (36.5)	60 (37.3)	90 (47.1)				
	other or unknown	7 (1.1)	2 (1.2)	3 (1.6)				
grade	g1	34 (5.4)	10 (6.2)	4 (2.1)	0.698	0.868	0.078	0.080
	g2	394 (62.5)	95 (59.0)	135 (70.7)				
	g3	162 (25.7)	41 (25.5)	41 (21.5)				
	g4	1 (0.2)	1 (0.6)	0 (0.0)				
	unknown	39 (6.2)	14 (8.7)	11 (5.8)				

Figure 3 – Structure learning constraints

	from	sex	age	grade	location	pT	pN	comorbidities	ASA	adj_therapy	recfree_01y	recfree_02y	surv_01y	surv_02y	surv_03y	surv_04y	surv_05y
to																	
sex		-	.	.	X	.	.	.	.	X	X	X	X	X	X	X	X
age		.	-	.	X	.	.	.	.	X	X	X	X	X	X	X	X
grade		.	.	-	.	X	X	.	.	.	X	X	X	X	X	X	X
location		.	.	.	-	.	.	.	.	X	X	X	X	X	X	X	X
pT		.	.	.	.	-	X	.	.	X	X	X	X	X	X	X	X
pN		.	.	.	.	.	-	.	.	X	X	X	X	X	X	X	X
comorbidities		.	.	.	.	.	.	-	X	X	.	.	.	.	.	.	.
ASA		.	.	.	.	.	.	.	-	X	X	X	X	X	X	X	X
adj_therapy		.	.	.	.	.	.	.	.	-	X	X	X	X	X	X	X
recfree_01y		.	.	.	.	.	.	.	.	.	-	X	X	X	X	X	X
recfree_02y		.	.	.	.	.	.	.	.	.	.	-	X	X	X	X	X
surv_01y		.	.	.	.	.	.	.	.	.	.	.	-	X	X	X	X
surv_02y		.	.	.	.	.	>	.	.	.	.	.	.	-	X	X	X
surv_03y		.	.	.	.	.	.	.	.	>	.	.	.	.	-	X	X
surv_04y		.	.	.	.	.	.	.	.	>	.	.	.	.	.	-	X
surv_05y		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-

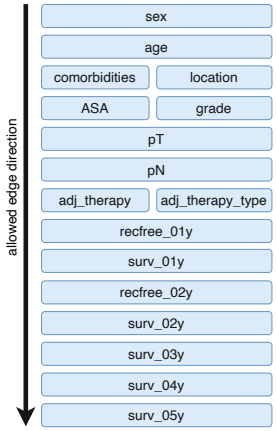


Figure 3: Structure learning constraints for the effect of adjuvant chemotherapy on overall survival. *Top:* Constraints for Hugin. The blacklist is made up from cells containing 'x': these edges are forbidden. Minimal whitelist (i.e., *enforced* edges) is formed by cells marked with '>'. *Bottom:* Constraints for bnstruct. Edges can only originate from a node in a layer at the same height or higher. For example, "sex" → "age" is allowed, but not the reverse. Edges between nodes at the same height can point either way.

Figure 4 – HRs

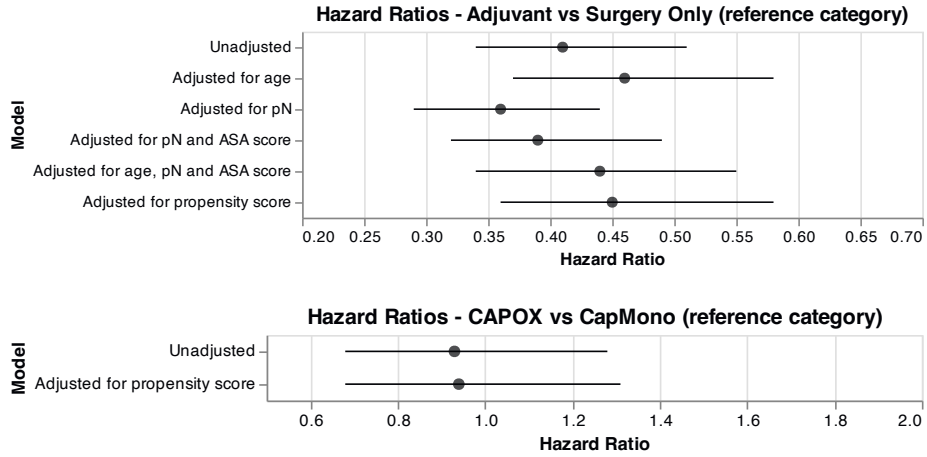


Figure 4: Hazard Ratio's (with confidence intervals). *Top:* adjuvant treatment vs surgery only (referent category). *Bottom:* CAPOX vs CapMono (referent category).

Figure 5 – BNs – Adjuvant/Surgery Only

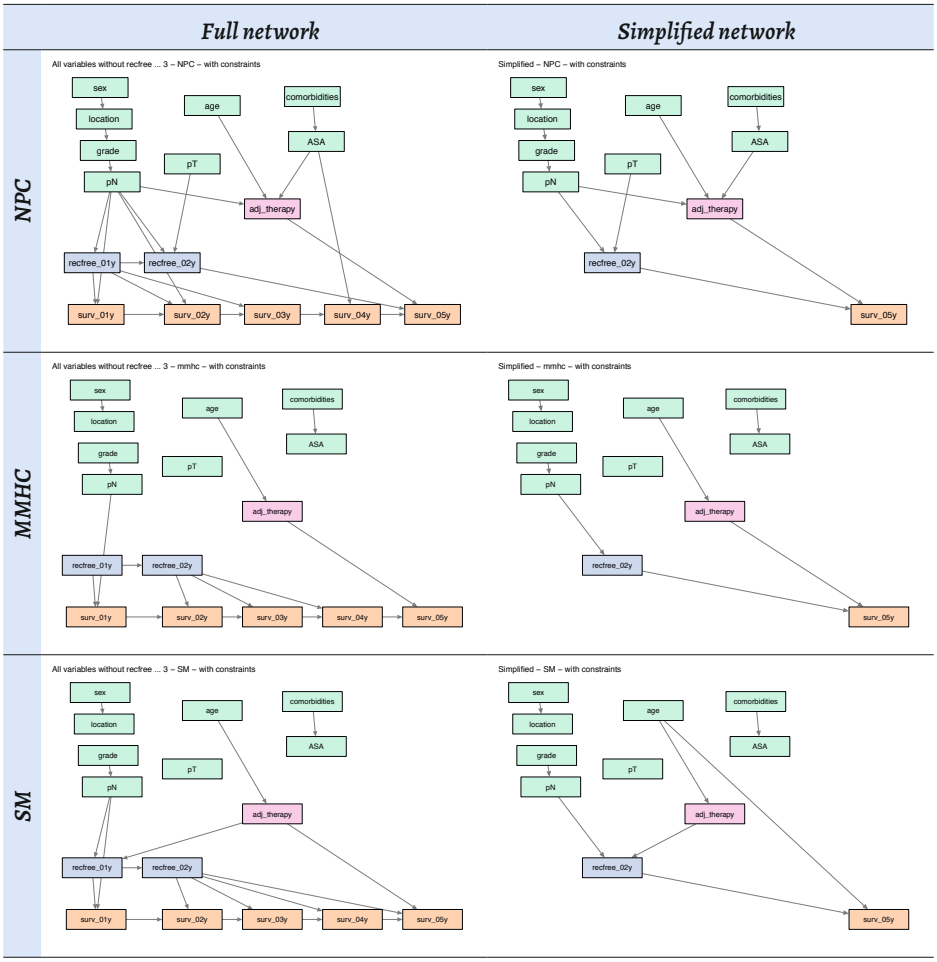


Figure 5: Bayesian Networks for adjuvant treatment vs surgery only. Rows 1 – 3 correspond to the algorithms NPC, MMHC. and SM respectively. Left column shows full networks, right column shows simplified networks. Treatment indicates whether any form of adjuvant treatment was received. *Top left:* pN and ASA act as confounders. *Top right:* pN acts as a confounder. *Middle left:* No confounders identified; pN, ASA, and treatment are independent. *Middle right:* No confounders identified. *Bottom left:* No confounders identified; pN, ASA, and treatment are independent. *Bottom right:* Age acts as a confounder.

Figure 6 – BNs – CAPOX/CapMono

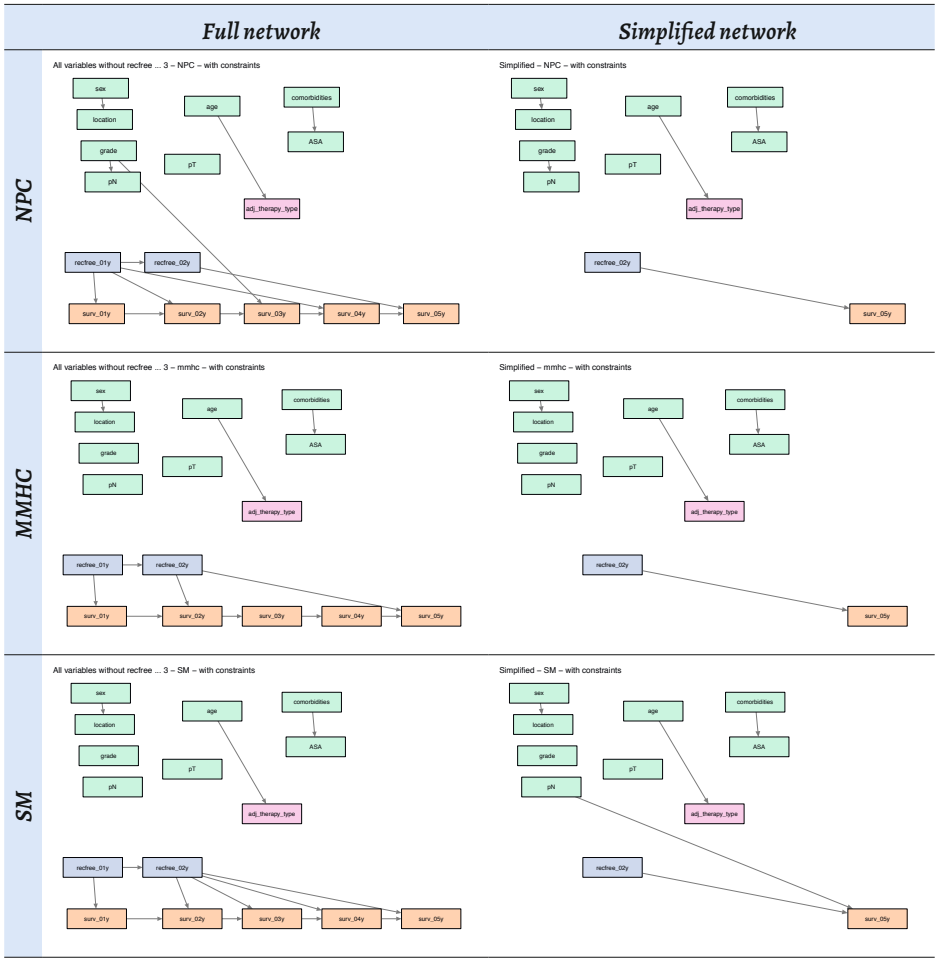


Figure 6: Bayesian Networks for CAPOX vs CapMono. Rows 1 – 3 correspond to the algorithms NPC, MMHC. and SM respectively. Left column shows full networks, right column shows simplified networks. Treatment indicates whether CAPOX or CapMono was received. *Top left:* No confounders found. *Top right:* No confounders found. *Middle left:* No confounders found. *Middle right:* No confounders found. *Bottom left:* No confounders found. *Bottom right:* No confounders found.

References

1. NKR Cijfers | Incidentie - Grafiek [Internet][cited 2023 Mar 2] Available from: https://nkr-cijfers.iknl.nl
2. Erning FN van, Janssen-Heijnen MLG, Creemers GJ, et al: Recurrence-free and overall survival among elderly stage III colon cancer patients treated with CAPOX or capecitabine monotherapy. Int J Cancer 140:224–233, 2017
3. Tournigand C, André T, Bonnetain F, et al: Adjuvant therapy with fluorouracil and oxaliplatin in stage II and elderly patients (between ages 70 and 75 years) with colon cancer: subgroup analyses of the Multicenter International Study of Oxaliplatin, Fluorouracil, and Leucovorin in the Adjuvant Treatment of Colon Cancer trial. J Clin Oncol Off J Am Soc Clin Oncol 30:3353–3360, 2012
4. Rubin DB, Thomas N: Matching using estimated propensity scores: relating theory to practice. Biometrics 52:249–264, 1996
5. Rubin DB: Estimating causal effects from large data sets using propensity scores. Ann Intern Med 127:757–763, 1997
6. Li J, Handorf E, Bekelman J, et al: Propensity score and doubly robust methods for estimating the effect of treatment on censored cost. Stat Med 35:1985–1999, 2016
7. Genbäck M, de Luna X: Causal inference accounting for unobserved confounding after outcome regression and doubly robust estimation. Biometrics 75:506–515, 2019
8. Mantel N, Haenszel W: Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. JNCI J Natl Cancer Inst 22:719–748, 1959
9. Mantel N: Chi-Square Tests with One Degree of Freedom; Extensions of the Mantel-Haenszel Procedure. J Am Stat Assoc 58:690–700, 1963
10. Häggström J: Data-driven confounder selection via Markov and Bayesian networks. Biometrics 74:389–398, 2018
11. Pearl J: Causality: Models, Reasoning and Inference (ed 2nd). New York, NY, USA, Cambridge University Press, 2009
12. VanderWeele TJ, Shpitser I: A new criterion for confounder selection. Biometrics 67:1406–1413, 2011
13. Sieswerda M, Xie S, van Rossum R, et al: Identifying Confounders Using Bayesian Networks and Estimating Treatment Effect in Prostate Cancer With Observational Data. JCO Clin Cancer Inform 7:e2200080, 2023
14. Koller D, Friedman N: Probabilistic Graphical Models: Principles and Techniques. Cambridge, Massachusets; London, England, MIT Press, 2009
15. Steck H: Constraint-based structural learning in Bayesian networks using finite data sets, in 2001
16. Silander T, Myllymaki P: A simple approach for finding the globally optimal Bayesian network structure [Internet], 2012[cited 2022 Nov 15] Available from: http://arxiv.org/abs/1206.6875
17. Tsamardinos I, Brown LE, Aliferis CF: The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn 65:31–78, 2006
18. Andersen SK, Olesen KG, Jensen FV, et al: HUGIN - a Shell for Building Bayesian Belief Universes for Expert Systems. Proc Elev Int Jt Conf Artif Intell 2:1080–1085, 1989
19. R Core Team: R: A language and environment for statistical computing [Internet]. Vienna, Austria, 2022Available from: https://www.R-project.org/
20. Franzin A, Sambo F, di Camillo B: bnstruct: an R package for Bayesian Network structure learning in the presence of missing data. Bioinforma Oxf Engl 33:1250–1252, 2017
21. Epskamp S, Cramer AOJ, Waldorp LJ, et al: qgraph: Network Visualizations of Relationships in Psychometric Data. J Stat Softw 48:1–18, 2012
22. Agresti A: Categorical Data Analysis Second Edition. Wiley, New York, 2002
23. DeCosse JJ, Ngoi SS, Jacobson JS, et al: Gender and colorectal cancer. Eur J Cancer Prev Off J Eur Cancer Prev Organ ECP 2:105–115, 1993

24. Marija C, Kresimir D, Ognjen B, et al: Estimation of colon cancer grade and metastatic lymph node involvement using DWI/ADC sequences. Acta Radiol Stockh Swed 1987 2841851221130008, 2022
25. Quach LH, Jayamaha S, Whitehouse SL, et al: Comparison of the Charlson Comorbidity Index with the ASA score for predicting 12-month mortality in acute hip fracture. Injury 51:1004–1010, 2020
26. Wilkinson NW, Yothers G, Lopa S, et al: Long-term survival results of surgery alone versus surgery plus 5-fluorouracil and leucovorin for stage II and stage III colon cancer: pooled analysis of NSABP C-01 through C-05. A baseline from which to compare modern adjuvant trials. Ann Surg Oncol 17:959–966, 2010
27. Haller DG, Tabernero J, Maroun J, et al: Capecitabine Plus Oxaliplatin Compared With Fluorouracil and Folinic Acid As Adjuvant Therapy for Stage III Colon Cancer. J Clin Oncol 29:1465–1471, 2011
28. André T, Boni C, Mounedji-Boudiaf L, et al: Oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment for colon cancer. N Engl J Med 350:2343–2351, 2004
29. André T, de Gramont A, Vernerey D, et al: Adjuvant Fluorouracil, Leucovorin, and Oxaliplatin in Stage II to III Colon Cancer: Updated 10-Year Survival and Outcomes According to BRAF Mutation and Mismatch Repair Status of the MOSAIC Study. J Clin Oncol Off J Am Soc Clin Oncol 33:4176–4187, 2015
30. Aparicio T, Francois E, Cristol-Dalstein L, et al: PRODIGE 34 – FFCD 1402 – ADAGE: Adjuvant chemotherapy in elderly patients with resected stage III colon cancer: A randomized phase 3 trial. Dig Liver Dis 48:206–207, 2016
31. Aparicio T, Bouché O, Etienne P-L, et al: Preliminary tolerance analysis of adjuvant chemotherapy in older patients after resection of stage III colon cancer from the PRODIGE 34-FFCD randomized trial. Dig Liver Dis Off J Ital Soc Gastroenterol Ital Assoc Study Liver S1590-8658(22)00662–4, 2022
32. Yothers G, O'Connell MJ, Allegra CJ, et al: Oxaliplatin as adjuvant therapy for colon cancer: updated results of NSABP C-07 trial, including survival and subset analyses. J Clin Oncol Off J Am Soc Clin Oncol 29:3768–3774, 2011
33. Schmoll H-J, Tabernero J, Maroun J, et al: Capecitabine Plus Oxaliplatin Compared With Fluorouracil/Folinic Acid As Adjuvant Therapy for Stage III Colon Cancer: Final Results of the NO16968 Randomized Controlled Phase III Trial. J Clin Oncol Off J Am Soc Clin Oncol 33:3733–3740, 2015

Supplemental Materials

Supplemental Tables & Figures

		Adjuvant therapy		p	sign.	p (CMH)	sign. (CMH)
		No (n=630)	Yes (n=352)				
variable	values						
sex	male	284 (45.1)	183 (52.0)	0.044	*	0.543	
	female	346 (54.9)	169 (48.0)				
age	70 - 74	108 (17.1)	192 (54.5)	0.000	***	0.126	
	75 - 79	197 (31.3)	131 (37.2)				
	80+	325 (51.6)	29 (8.2)				
comorbidities	none	97 (15.4)	98 (27.8)	0.000	***	0.943	
	1	135 (21.4)	100 (28.4)				
	2+	377 (59.8)	146 (41.5)				
	unknown	21 (3.3)	8 (2.3)				
ASA	1	16 (2.5)	37 (10.5)	0.000	***	0.516	
	2	241 (38.3)	189 (53.7)				
	3	220 (34.9)	47 (13.4)				
	4	4 (0.6)	1 (0.3)				
	unknown	149 (23.7)	78 (22.2)				
pT	T1	10 (1.6)	10 (2.8)	0.259		0.166	
	T2	51 (8.1)	36 (10.2)				
	T3	452 (71.7)	251 (71.3)				
	T4	117 (18.6)	55 (15.6)				
pN	N1	478 (75.9)	231 (65.6)	0.001	**	0.705	
	N2	152 (24.1)	121 (34.4)				
location	proximal	393 (62.4)	197 (56.0)	0.142		0.999	
	distal	230 (36.5)	150 (42.6)				
	other/unknown	7 (1.1)	5 (1.4)				
grade	g1	34 (5.4)	14 (4.0)	0.698		0.838	
	g2	394 (62.5)	230 (65.3)				
	g3	162 (25.7)	82 (23.3)				
	g4	1 (0.2)	1 (0.3)				
	unknown	39 (6.2)	25 (7.1)				

Table S1: Variable distribution - Adjuvant vs Surgery Only. Propensity scores were calculated using *sex*, *age*, *comorbidities*, *ASA*, and *pN*. 'p': calculated using Pearson Chi^2-test. 'p*' calculated using the Cochran-Mantel-Haenszel Chi^2-test after stratification by propensity score quintile. *: $p \leq 0.05$, **: $p \leq 0.005$, ***: $p \leq 0.0005$.

		Adjuvant therapy type		p	sign.	p (CMH)	sign. (CMH)
		CapMono (n=161)	CAPOX (n=191)				
variable	values						
sex	male	75 (46.6)	108 (56.5)	0.079		0.094	
	female	86 (53.4)	83 (43.5)				
age	70 - 74	52 (32.3)	140 (73.3)	0.000	***	0.129	
	75 - 79	84 (52.2)	47 (24.6)				
	80+	25 (15.5)	4 (2.1)				
comorbidities	none	33 (20.5)	65 (34.0)	0.043	*	0.291	
	1	52 (32.3)	48 (25.1)				
	2+	72 (44.7)	74 (38.7)				
	unknown	4 (2.5)	4 (2.1)				
ASA	1	12 (7.5)	25 (13.1)	0.141		0.220	
	2	83 (51.6)	106 (55.5)				
	3	27 (16.8)	20 (10.5)				
	4	0 (0.0)	1 (0.5)				
	unknown	39 (24.2)	39 (20.4)				
pT	T1	6 (3.7)	4 (2.1)	0.286		0.400	
	T2	18 (11.2)	18 (9.4)				
	T3	107 (66.5)	144 (75.4)				
	T4	30 (18.6)	25 (13.1)				
pN	N1	107 (66.5)	124 (64.9)	0.849		0.351	
	N2	54 (33.5)	67 (35.1)				
location	proximal	99 (61.5)	98 (51.3)	0.159		0.428	
	distal	60 (37.3)	90 (47.1)				
	other/unknown	2 (1.2)	3 (1.6)				
grade	g1	10 (6.2)	4 (2.1)	0.078		0.080	
	g2	95 (59.0)	135 (70.7)				
	g3	41 (25.5)	41 (21.5)				
	g4	1 (0.6)	0 (0.0)				
	unknown	14 (8.7)	11 (5.8)				

Table S2: Variable distribution - CAPOX vs CapMono. Propensity scores were calculated using *sex* and *comorbidities*. 'p': calculated using Pearson Chi^2-test. 'p*' calculated using the Cochran-Mantel-Haenszel Chi^2-test after stratification by propensity score quintile. *: $p \leq 0.05$, **: $p \leq 0.005$, ***: $p \leq 0.0005$.

Model	HR	95% CI	p	sign.
Unadjusted	0.41	0.34-0.51	0.0000	***
Adjusted for pN	0.36	0.29-0.44	0.0000	***
Adjusted for pN and ASA score	0.39	0.32-0.49	0.0000	***
Adjusted for age	0.46	0.37-0.58	0.0000	***
Adjusted for age, pN and ASA score	0.44	0.34-0.55	0.0000	***
Adjusted for propensity score	0.45	0.36-0.58	0.0000	***

Table S3: Hazard Ratios and Confidence Intervals for Adjuvant Treatment vs Surgery Only (referent category).

Model	HR	95% CI	p	sign.
Unadjusted	0.93	0.68-1.28	0.6669	
Adjusted for propensity score	0.94	0.68-1.31	0.7274	

Table S4: Hazard Ratios and Confidence Intervals for CAPOX vs CapMono (referent category).

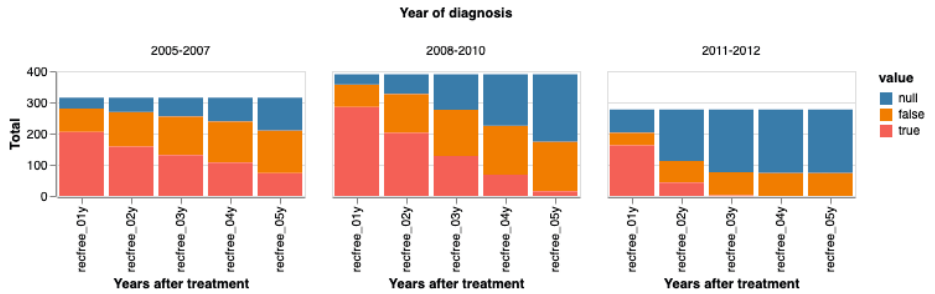


Figure S1: Recurrence-free survival, stratified by (categorized) year of diagnosis

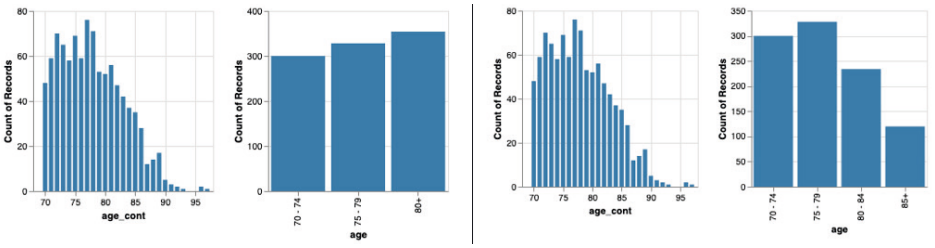


Figure S2: Effect of discretization of age into 3 or 4 categories.

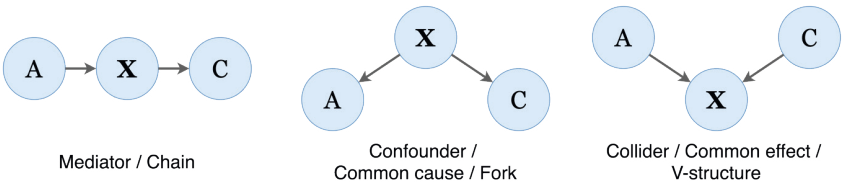


Figure S3: Building blocks of Bayesian Networks

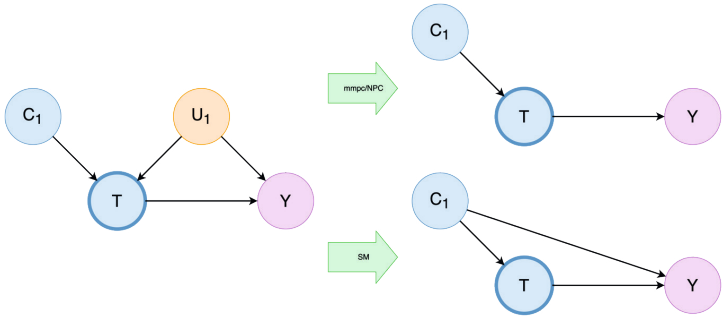


Figure S4: Reconstruction of a model with an unmeasured confounder. The network on the left represents the ground truth and is used to generate a dataset. Here C_1 is a pre-treatment variable, T represents treatment and Y represents outcome. U_1 is an unmeasured confounder. The right models show the graphs identified when applying structure learning to this data with column "U_1" removed, using either MMPC or NPC (top) or SM (bottom). SM incorrectly adds the edge $C_1 \rightarrow Y$. Note that this specific outcome may depend on the underlying CPTs/associations, and it may be possible to construct CPTs that would also cause MMPC/NPC to inappropriately edge $C_1 \rightarrow Y$.

Variable selection

After a brief investigation into the association of year of diagnosis with recurrence free survival (Figure 3), we decided to ignore the variable in our analyses. In the later years there were too many missing values as a result of limited follow-up and this caused spurious associations to be found.

Discretization of age

The choice for three categories was ultimately arbitrary, although a few thoughts/reasons drove the decision:

- We thought five-year intervals are frequently used in medical literature and in communication to patients and are easy to interpret.
- Using three bins gave a reasonably uniform distribution. Adding a bin would make the category 85+ about a third of the category 75-79.
- The number of intervals is directly related to the (mathematical) complexity of the model. For example, in cases of simple, direct confounding (survival ← age → treatment, treatment → survival) adding a single bin would add 6 parameters to these variables alone. More is not necessarily better, especially given the limited amount of data.

7

Predicting treatment effect on patient-reported outcomes and survival in rectal cancer

Adapted from Melle S. Sieswerda, Maaike Berbee, Martijn Intven, Inigo Bermejo, Gijs Geleijnse, Felice van Erning, Iris Walraven, Valery Lemmens, André Dekker, Xander Verbeek; "**Predicting treatment effect on patient-reported outcomes and survival in rectal cancer**"; In preparation for submission to *JCO Clin Cancer Inform*.

Abstract

In the Netherlands, rectal cancer is the fifth most frequently observed malignancy. Treatment options for nonmetastatic disease depend on stage but generally consist of surgery with or without neoadjuvant radiotherapy or chemoradiation. While neoadjuvant treatment improves local control and enables better curative options, it also incurs the risk of overtreatment and inferior functional outcome in some patients. The use of clinical decision support (CDS) systems has the potential to guide personalized care and shared decision making (SDM) by providing meaningful prognostic information. Existing models are not suitable to guide personalized care.

Bayesian Networks (BNs) are a type of probabilistic graphical model that can be used to estimate the probability distribution of a variable given evidence and can be learned from data while incorporating existing knowledge. In contrast to other models, they do not have dedicated inputs or outputs. Instead, applying evidence updates probabilities throughout the network. This, in combination with their visual representation and ability for causal analyses, makes them well suited for supporting SDM.

As a first step towards supporting effective shared decision making surrounding neoadjuvant (chemo)radiotherapy in patients with nonmetastatic rectal cancer, we hypothesized BNs can be used to simultaneously predict recurrence-free survival, overall survival, and quality of life given treatment choice for patients with nonmetastatic rectal cancer.

Contrary to our expectations, however, structure learning initially did not identify any associations between treatment and survival, recurrence, or PROMs. However, associations between treatment and survival *were* found if recurrence-free survival nodes were excluded from the model. By manually adding relationships to the BN we were still able to simultaneously predict survival and PROMs with reasonable calibration.

Introduction

In the Netherlands, rectal cancer is the fifth most frequently observed malignancy, with an incidence around 4700 new cases per year.[1] Treatment options include surgery only, neoadjuvant radiotherapy (short-course), and neoadjuvant chemoradiation (long-course). Neoadjuvant radiotherapy has the primary benefit of improving local control (compared to surgery alone); in more (locally) advanced stage, neoadjuvant chemoradiation additionally improves survival and the probability of sphincter preservation compared to surgery with or without radiotherapy.[2–5] However, (neo)adjuvant treatment also has potential downsides, such as toxicity, increased risk of perianal complications after surgery, or reduced functional outcomes such as fecal incontinence.[6]

In the Netherlands, the choice for neoadjuvant therapy is a shared decision between patient and radiation oncologist. Although through shared decision making (SDM), a balance is sought between the risks and benefits of a particular treatment, currently this decision is largely based on the risk of local recurrence. It is felt that the decision-making process would benefit from the ability to incorporate patient reported outcome measures (PROMs) and health-related Quality of Life (HRQoL). The use of clinical decision support (CDS) systems with personalized risk prediction models has the potential to guide personalized care by providing physicians and patients with meaningful and objective prognostic information and help strike a balance between QoL and treatment side effects.[7]

Currently, predictive models for rectal cancer exist but mostly focus on either survival, recurrence or late side effects.[8–13] However, all these types of outcomes are important when it comes to the choice whether to start treatment. Consequently, existing models are not optimal to guide personalized care and support decision making. As such, it would be helpful to develop a model that simultaneously predicts recurrence, survival, and PROMs based on treatment.

Bayesian Networks (BNs) are a type of probabilistic graphical model that use a directed acyclic graph (DAG). The nodes represent variables, and the directed edges signify conditional dependencies, preferably causal relationships. A node with an edge pointing towards another node, is referred to as the other node's parent. Each node is associated with a probability distribution that is conditional on its parents. Relationships between three nodes can be broken down into three building blocks, named after the role of the "middle" variable: mediators ($X \rightarrow Y \rightarrow Z$), common causes or confounders ($X \leftarrow Y \rightarrow Z$), and V-structures ($X \rightarrow Y \leftarrow Z$). For mediators and common causes, X and Z are not independent (denoted as $X \not\perp\!\!\!\perp Z$), but X is *conditionally* independent of Z given Y (denoted as $X \perp\!\!\!\perp Z \mid Y$). For V-structures the situation is reversed: X and Z are independent ($X \perp\!\!\!\perp Z$), but X and Z are conditionally dependent given Y ($X \not\perp\!\!\!\perp Z \mid Y$). Consequently, conditional (in)dependencies can be read of the graph. Additionally, BNs can be used to estimate the probability distribution of a variable given evidence. In contrast to other

models, such as logistic regression or neural networks, BNs do not have dedicated inputs or outputs. Instead, setting evidence on any node updates probabilities throughout the network. This, in combination with their visual representation, makes them exceptionally suited for supporting SDM.

As a first step towards supporting effective SDM surrounding neoadjuvant (chemo) radiotherapy in patients with rectal cancer, we hypothesized BNs can be used to simultaneously predict local recurrence, overall survival, and HRQoL given treatment choice for patients with nonmetastatic rectal cancer.

Methods

Data

Patients diagnosed with rectal carcinoma between 2015 and 2019 and with age between 40 and 90 years old at time of diagnosis were selected from the Netherlands Cancer Registry (NCR). Patients with metastasized disease at time of diagnosis (cM1) were excluded as were patients that underwent endoscopic resection, primary chemotherapy, or received no treatment at all. Vital status was updated until January 31st, 2024.

Variables describing gender, age at diagnosis, involvement of the mesorectal fascia (MRF), clinical tumor stage (cTNM), vital status, and date of last follow-up were acquired. Additionally, details regarding (neoadjuvant) treatment were obtained.

In the period of interest (2015 – 2019), Dutch guidelines recommended, depending on tumor stage and risk factors, either short course radiotherapy (5x5 Gy) or long course chemoradiation (25x2 Gy or 28x1.8 Gy, and capecitabine). Therefore, patients that received deviating treatment were excluded (n=356).

Patients that received (chemo)radiotherapy only were assumed to have been treated with neoadjuvant intent. Consequently, treatment was grouped into neoadjuvant radiotherapy, neoadjuvant chemoradiation, and surgery only. For analysis in the BN, treatment was replaced with the variable "neoadjuvant" (with values none, radiotherapy, and chemoradiation), and the binary variable surgery was derived to indicate whether surgery took place.

For assessing association with treatment and for use in the BN, age was discretized into 5 categories: 41 – 50, 51 – 60, 61 – 70, 71 – 80, and 81 – 90.

In the NCR, MRF involvement is generally only registered for cT3 tumors. However, for other stages it may be registered if reported on diagnostic imaging. As MRF is only defined for tumors ≥ cT3, the MRF for patients with tumors ≤ cT2 was set to "not

applicable". If necessary, involvement of the MRF was derived based on the distance between the tumor and the MRF. Here, a distance > 1mm was considered negative and, vice versa, ≤ 1mm was considered positive.

Overall survival (OS), counted from date of diagnosis, was discretized into five Boolean variables representing 1- to 5-year survival, introducing NAs (missing values) for patients who had a follow-up < 5 years and were alive at the time of follow-up (Supplemental Materials - Figure S1). For a subset of patients (n=1477), diagnosed between January 1st and June 30th 2015, information about recurrence was available. Absence of recurrence was discretized similarly as survival, yielding variables for recurrence-free survival (RFS). Due to the limited follow-up, variables representing recurrence ≥ 4y were skewed toward recurrence or NA (Supplemental Materials - Figure S2).

The patient selection was linked to data from the Prospective Dutch CRC cohort (PLCRC). For 584 patients responses to several PROMs-questionnaires were obtained, specifically the EORTC QLQ-C30, and the Low Anterior Resection Syndrome (LARS) score.[14] Responses were available at multiple timepoints, corresponding roughly to 0, 3, 6, 12, 18, 24, 36, 48, 60, and 72 months after diagnosis, although data completeness deteriorated with passing of time.

Response dates were used to identify baseline response (first response before treatment) and effect (first response 1 year after diagnosis). The individual answers to the C30 questionnaire were used to calculate domain scores that were subsequently dichotomized (i.e., as clinically important or unimportant) as described in Giesinger et al.[15] Only domain scores related to symptoms were kept, i.e. fatigue (FA), nausea/vomiting (NV), pain (PA), dyspnea (DY), sleep disturbances (SL), appetite loss (AP), constipation (CO), and diarrhea (DI). Additionally, the discretized interpretation of LARS-score (i.e., no, minor, or major LARS) was selected.

Statistical analyses & Bayesian Network development

Association between pre-treatment variables and treatment was evaluated using the Chi^2 test. Association between PROMs at baseline and effect was analyzed similarly.

Bayesian Network development was performed as described in Sieswerda et al.[16,17] Briefly, to obtain a causal model, structure learning was performed using the PC-algorithm as implemented in Hugin, version 7.4, together with a blacklist containing known non-causal edges. Level of significance was set to 0.05. Where absent (i.e., not found by structure learning), edges between treatment and recurrence, survival, and PROMs were manually added through a whitelist to obtain a final model.

The Expectation Maximization (EM) algorithm was subsequently used for parameter estimation.[18] Structure learning was performed on the full dataset. For parameter

estimation and evaluation of model performance, the dataset was split into a training and test set consisting of 60% (n=8273) and 40% (n=5516) of the full dataset respectively.

Model evaluation

Since the goal of the model was a) to predict both PROMs *and* survival simultaneously, and b) to be used clinically, it was decided to focus primarily on model calibration; the model's ability to accurately predict a specific outcome was considered less important than its ability to correctly predict a probability. Calibration plots were therefore created as follows. Using the test-set, for each outcome node (i.e. nodes for RFS-, OS-, and effect-PROMs), sizes of subgroups based on treatment (i.e. variables neoadjuvant and surgery) were determined and for each subgroup the predicted probability (x-axis) was plotted against the observed frequency (y-axis). Additionally, calibration for the treatment variables, neoadjuvant and surgery, was calculated using their respective parents to determine the subgroups.

Results

Variable distribution and Chi² analysis

The final dataset contained 13,789 patients of which 64% was male and 36% female. Average age at diagnosis was 67.5 years (SD: 10.1), median age was 68 years. The majority of patients were clinically staged as 1, 2A, or 3B (n=3523, n=2695, and n=4832 respectively). Very few patients were staged as 2B (n=98) or 2C (n=183). Patients that underwent surgery only, mostly had a tumor stage 1 or 2A. Neoadjuvant radiotherapy was more frequently given in stages 3A and 3B, while neoadjuvant chemoradiation was more frequent in stages 3B and 3C. Variable distribution is shown in Table 1; full variable distribution is available in Supplemental Materials – Table S1. Distribution of tumor stage is visualized in Supplemental Materials – Figure S3. Of the patients that underwent neoadjuvant radiotherapy (n=3382), 2760 (82%) proceeded to surgery. For those that received neoadjuvant chemoradiation (n=4632) 3562 (77%) subsequently had surgery.

The variables age, cTNM, 1- to 5-year RFS, 1- to 5-year OS, fatigue after treatment, pain at baseline, and diarrhea at baseline were significantly associated with treatment at $p \leq 0.05$.

Distribution of the PROMs was visualized (Supplemental Materials – Figure S4). The distributions of all PROMs domains were skewed towards being clinically important, with diarrhea (DI) being the least pronounced. Association between baseline and effect, as determined using the Chi² test, was significant in 7 out of 9 PROMs.

Bayesian Network

A blacklist with known non-causal edges was created (Table 2) and used during structure learning to develop an initial BN (Figure 1, top-left panel). No associations were found between treatment (i.e., nodes neoadjuvant and surgery) and survival; neither did structure learning find associations between treatment and recurrence, nor between treatment and PROMs. The algorithm did find an association between cTNM-stage and neoadjuvant treatment, as well as edges between node MRF and nodes cTNM, neoadjuvant, and treatment.

With respect to recurrence and survival, structure learning identified several associations. Each recurrence node was predictive of its subsequent time point (e.g., 1-year RFS predicted 2-year RFS, and so forth). Additionally, 1- and 2-year RFS predicted 1- and 2-year OS respectively, and 3- and 4-year RFS respectively predicted 4- and 5-year OS.

With respect to the PROMs nodes, only two baseline PROMs were connected to their corresponding effect nodes: fatigue (FA) and dyspnea (DY). Additionally, an edge was found between constipation (CO) at baseline and appetite loss (AP) after treatment. Furthermore, two clusters could be identified: one within the baseline PROMs and one within the effect PROMs. One edge was found between sex and sleep disturbances (SL) after treatment.

The cluster in the baseline PROMs consisted of the nodes for fatigue (FA), pain (PA), sleep disturbances (SL), appetite loss (AP), and nausea/vomiting (NV). In this cluster, pain and appetite loss were conditionally independent given fatigue ($PA \perp\!\!\!\perp AP | FA$). In other words, pain and appetite loss are independent only if the presence or absence of fatigue is known and we stratify by fatigue. Fatigue and nausea/vomiting were conditionally dependent on appetite loss ($FA \not\perp\!\!\!\perp NV | AP$), indicating they are correlated if stratified by appetite loss, but uncorrelated otherwise. Similarly, fatigue and sleep disturbances were conditionally dependent on pain ($FA \not\perp\!\!\!\perp SL | PA$).

Within the effect PROMs a smaller cluster consisted of fatigue (FA), nausea/vomiting (NV), diarrhea (DI), and LARS. Here, diarrhea and fatigue were conditionally dependent on nausea/vomiting ($FA \not\perp\!\!\!\perp DI | NV$), and LARS and nausea/vomiting were conditionally independent on diarrhea ($LARS \perp\!\!\!\perp NV | DI$).

In order to elucidate the lack of association found between treatment and survival, and treatment and PROMs, structure learning was repeated three more times while excluding different sets of nodes: 1) excluding RFS- and PROMs-nodes, 2) excluding PROMs-nodes, and 3) excluding RFS- and OS-nodes (Figure 1, top-right, bottom-right, and bottom-left panel respectively). When RFS- and PROMs-nodes were excluded, structure learning was able to pick up associations between treatment (nodes neoadjuvant and surgery) and

survival nodes. These associations mostly disappeared when RFS was re-added (with the exception of the edge surgery → 2-year survival). Excluding RFS- and OS-nodes did not reveal new associations between treatment and PROMs.

To better understand the discrepancy between the Chi^2-analyses, which showed significant associations in 7 out of 9 PROMs between baseline and effect, and the limited number of associations found through structure learning, the observations were cross-tabulated (Figure 2). This highlighted that a minority of patients had PROMs measurements available at two time points, which provided an explanation for the discrepancy: structure learning uses conditional probabilities, which require data at two time points, where the Chi^2-test merely compares distributions.

Finally, edges between treatment and outcome nodes were manually added, together with the edges sex → cTNM, and sex → MRF which were found through structure learning while excluding recurrence-free survival and PROMs node. EM-learning was performed using the training dataset. The final network is shown in Figure 3 and Figure 4; Hugin OOBN and NET files are available as supplemental files.

The final network was used to create the calibration plots shown in Figure 5. 1- to 3-year RFS appeared reasonably calibrated for patients that underwent surgery (with or without neoadjuvant therapy). Furthermore, 1- to 5-year survival was well calibrated when using the treatment nodes as input: every subgroup was (very nearly) on the diagonal.

Calibration for the PROMs nodes varied, with appetite loss (AP), constipation (CO), nausea and vomiting (NV), and sleep disturbances (SL) being reasonably calibrated, but diarrhea (DI), dyspnea (DY), fatigue (FA), pain (PA), and LARS less so.

Prediction of neoadjuvant treatment, which had (3*(2*3*8) =) 144 subgroups because of its three parents, was well calibrated for the larger subgroups, but less so for smaller subgroups; prediction given cTNM only, resulting in 24 subgroups, was well calibrated and is shown in Supplemental Materials – Figure S5. Prediction of surgery was similarly calibrated. Here too, calculating calibration using a fewer number of subgroups improved calibration, as shown in Supplemental Materials – Figure S6.

When inspecting the prior probabilities for the recurrence-free survival and overall survival nodes, probability of a favorable outcome decreased over the years. Prior probability of 5-year survival was estimated at 74.5%. Probabilities of 5-year survival given patients underwent surgery only, neoadjuvant radiotherapy, or neoadjuvant chemoradiation, were 80.7%, 75.5%, and 77.1% respectively.

Discussion & Conclusion

In this study we developed a Bayesian Network that simultaneously predicts recurrence, survival, and PROMs as a basis for shared decision making in rectal cancer. While the model was well calibrated for survival, calibration for 1- to 3-year recurrence-free survival was only acceptable for the patients that underwent surgery, and only 4 out of 9 PROMs were reasonably calibrated. For the final model, edges between treatment and outcome nodes had to be manually added because structure learning did not find associations between treatment and PROMs nodes, nor did it find associations between treatment and (recurrence-free) survival. However, associations between treatment and survival *were* found if recurrence-free survival nodes were excluded from the model.

The lack of association between treatment and recurrence-free survival might be explained by insufficient data, both in terms of number of patients (n=1477) and duration of follow-up (n=684 for 4-year recurrence-free survival). A true lack of association seems less plausible, considering associations between treatment and (overall) survival were found, and recurrence-free survival was in turn associated to overall survival.

Similarly, the lack of association between treatment and PROMs, and PROMs at baseline and effect may be the result of poor data-availability. This seems especially plausible for the latter associations, since a minority of patients had PROMs measurements available at two time points. Still, true lack of association cannot be ruled out completely: in the calibration plots, the different subgroups are poorly separated. One possible explanation may be found in response shift where measured effect is dampened because patients adapt to their new circumstances and outlook.[19, 20] Further research is required to better understand the dynamics at play.

When comparing predictions for 5-year survival with literature, our model predicted an a-priori probability of 74.5%, which is higher than the 62% estimated by van Gijn et al. in 2011.[2] However, prognosis has improved in recent years. For example, statistics from the Netherlands Comprehensive Cancer Organisation estimate 5-year survival for patients diagnosed with rectal cancer between 2015 – 2022 at 70%.[21] The remaining difference may be explained by our exclusion of metastasized disease, although we also excluded patients that underwent endoscopic resection. For patients that underwent surgery only, 3-year recurrence-free survival was predicted at 85.7%, which appears in line with a paper from Heald and Ryall, where Figure 1 suggests that 3-year recurrence-free survival was around 87%.[22]

Looking at the network structure in general, node MRF is a parent to nodes neoadjuvant, cTNM, and surgery. While this seems clinically plausible, the association with cTNM may be an artifact introduced by how MRF was discretized that caused cT1-2 tumors to be strongly associated with MRF state “Not Applicable”. Additionally, cTNM did not appear to be independently associated with survival; any effect was mediated through

nodes neoadjuvant and surgery. While this may seem counterintuitive at first, it may in fact be the effect of good guideline adherence.

Inspecting the associations between the PROMs nodes revealed two clusters. The associations in the cluster between the baseline PROMs can be intuitively explained, with the exception perhaps the V-structure fatigue → pain ← sleep disturbances. This suggests fatigue and sleep disturbances can independently predict pain, and fatigue and sleep disturbances are conditionally dependent. In other words: fatigue is not necessarily caused by sleep disturbances and vice versa. In the effect cluster the association diarrhea → nausea/vomiting is intuitive, although the direction of influence seems more likely to be reversed. This might have been avoided if the blacklist had contained associations pertaining to interactions between PROMs.

In conclusion, we developed a model that can simultaneously predict survival and PROMs in patients with rectal cancer and is reasonably calibrated. However, given the limited separation between PROMs-outcomes across treatments, the usefulness of the model appears limited.

Tables & Figures

Tables

variable	values	neoadj. chemoradiation (n=4632)	neoadj. radiotherapy (n=3382)	surgery only (n=5775)	p	sign.
age	**41 - 50**	437 (9.4)	117 (3.5)	183 (3.2)	0.0000	***
	51 - 60	1222 (26.4)	552 (16.3)	937 (16.2)		
	61 - 70	1600 (34.5)	971 (28.7)	2188 (37.9)		
	71 - 80	1198 (25.9)	971 (28.7)	1869 (32.4)		
	81 - 90	175 (3.8)	771 (22.8)	598 (10.4)		
sex	**male**	2887 (62.3)	2138 (63.2)	3741 (64.8)	0.0315	*
	female	1745 (37.7)	1244 (36.8)	2034 (35.2)		
MRF	**Not Applicable**	861 (19.9)	893 (28.5)	3892 (70.3)	0.0000	***
	MRF-	1093 (25.3)	1563 (49.9)	1388 (25.1)		
	MRF+	2362 (54.7)	678 (21.6)	253 (4.6)		
cTNM	**1**	126 (2.7)	160 (4.7)	3237 (56.1)	0.0000	***
	2A	601 (13.0)	604 (17.9)	1490 (25.8)		
	2B	58 (1.3)	28 (0.8)	12 (0.2)		
	2C	114 (2.5)	54 (1.6)	15 (0.3)		
	3A	111 (2.4)	562 (16.6)	133 (2.3)		
	3B	2784 (60.1)	1787 (52.8)	261 (4.5)		
	3C	826 (17.8)	129 (3.8)	24 (0.4)		
	X	12 (0.3)	58 (1.7)	603 (10.4)		
PROMs Available	**FALSE**	4379 (94.5)	3254 (96.2)	5572 (96.5)	0.0000	***
	TRUE	253 (5.5)	128 (3.8)	203 (3.5)		
recfree_04y	**FALSE**	164 (62.1)	87 (52.7)	104 (40.8)	0.0000	***
	TRUE	100 (37.9)	78 (47.3)	151 (59.2)		
surv_05y	**FALSE**	1096 (26.7)	1208 (38.9)	1091 (21.4)	0.0000	***
	TRUE	3013 (73.3)	1900 (61.1)	4005 (78.6)		

Table 1: Variable distributions and association with treatment for primary variables as calculated using the Chi^2-test. The full table is available in *Supplemental Materials – Table 1*. The numbers between brackets denote percentages. *: $p \leq 0.05$, **: $p \leq 0.005$, ***: $p \leq 0.0005$. MRF: mesorectal fascia; cTNM: clinical TNM-stage group, recfree_04y: 4-year recurrence-free survival; surv_05y: 5-year survival.

Destination \ Source	age_cat	sex	cTNM	MRF	neoadjuvant	surgery	surv_01y	surv_02y	surv_03y	surv_04y	surv_05y	recfree_01y	recfree_02y	recfree_03y	recfree_04y	FA_00y
age_cat	-	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
sex	X	-	X	X	X	X	X	X	X	X	X	X	X	X	X	X
cTNM	.	>>	-	.	X	X	X	X	X	X	X	X	X	X	X	X
MRF	.	>>	X	-	X	X	X	X	X	X	X	X	X	X	X	.
neoadjuvant	.	.	.	.	-	X	X	X	X	X	X	X	X	X	X	.
surgery	.	.	>>	.	.	-	X	X	X	X	X	X	X	X	X	.
surv_01y	.	.	.	.	>>	>>	-	X	X	X	X	.	X	X	X	.
surv_02y	.	.	.	.	>>	>>	.	-	X	X	X	.	.	X	X	.
surv_03y	.	.	.	.	>>	>>	.	.	-	X	X	.	.	.	X	.
surv_04y	.	.	.	.	>>	>>	.	.	.	-	X	.	.	.	.	.
surv_05y	.	.	.	.	>>	>>	.	.	.	.	-	.	.	.	.	.
recfree_01y	.	.	.	.	>>	>>	X	X	X	X	X	-	X	X	X	.
recfree_02y	.	.	.	.	>>	>>	X	X	X	X	X	.	-	X	X	.
recfree_03y	.	.	.	.	>>	>>	X	X	X	X	X	.	.	-	X	.
recfree_04y	.	.	.	.	>>	>>	X	X	X	X	X	.	.	.	-	.
FA_00y	.	.	.	.	X	X	X	X	X	X	X	X	X	X	X	-
FA_01y	.	.	.	.	>>	>>	X	X	X	X	X	.	X	X	X	.
NV_00y	.	.	.	.	X	X	X	X	X	X	X	X	X	X	X	.
NV_01y	.	.	.	.	>>	>>	X	X	X	X	X	.	X	X	X	.
PA_00y	.	.	.	.	X	X	X	X	X	X	X	X	X	X	X	.
PA_01y	.	.	.	.	>>	>>	X	X	X	X	X	.	X	X	X	.
DY_00y	.	.	.	.	X	X	X	X	X	X	X	X	X	X	X	.
DY_01y	.	.	.	.	>>	>>	X	X	X	X	X	.	X	X	X	.
SL_00y	.	.	.	.	X	X	X	X	X	X	X	X	X	X	X	.
SL_01y	.	.	.	.	>>	>>	X	X	X	X	X	.	X	X	X	.
AP_00y	.	.	.	.	X	X	X	X	X	X	X	X	X	X	X	.
AP_01y	.	.	.	.	>>	>>	X	X	X	X	X	.	X	X	X	.
CO_00y	.	.	.	.	X	X	X	X	X	X	X	X	X	X	X	.
CO_01y	.	.	.	.	>>	>>	X	X	X	X	X	.	X	X	X	.
DI_00y	.	.	.	.	X	X	X	X	X	X	X	X	X	X	X	.
DI_01y	.	.	.	.	>>	>>	X	X	X	X	X	.	X	X	X	.
LARS_01y	.	.	.	.	>>	>>	X	X	X	X	X	.	X	X	X	.

Table 2: Structure Learning constraints. The blacklist is made up from cells containing 'x': these edges are forbidden. The whitelist (i.e., *manually added* edges) is formed by cells marked with '>>'. MRF: mesorectal fascia; cTNM: clinical TNM-stage group, recfree: recurrence-free survival; surv: survival. FA: Fatigue; NV: Nausea/Vomiting; PA: Pain, DY: Dyspnea, SL: Sleep Disturbances; AP: Appetite Loss; CO: constipation; DI: Diarrhea; LARS: Low Anterior Resection Syndrome score. The suffix _0{n}y for recurrence, survival, and PROMs, indicates the number of years after diagnosis.

FA_01y	NV_00y	NV_01y	PA_00y	PA_01y	DY_00y	DY_01y	SL_00y	SL_01y	AP_00y	AP_01y	CO_00y	CO_01y	DI_00y	DI_01y	LARS_01y
X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
X	.	X	.	X	.	X	.	X	.	X	.	X	.	X	X
X	.	X	.	X	.	X	.	X	.	X	.	X	.	X	X
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
X	.	X	.	X	.	X	.	X	.	X	.	X	.	X	X
-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X
X	-	X	.	X	.	X	.	X	.	X	.	X	.	X	X
.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	X
X	.	X	-	X	.	X	.	X	.	X	.	X	.	X	X
.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	X
X	.	X	.	X	-	X	.	X	.	X	.	X	.	X	X
.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	X
X	.	X	.	X	.	X	-	X	.	X	.	X	.	X	X
.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	X
X	.	X	.	X	.	X	.	X	-	X	.	X	.	X	X
.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	X
X	.	X	.	X	.	X	.	X	.	X	-	X	.	X	X
.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	X
X	.	X	.	X	.	X	.	X	.	X	.	X	-	X	X
.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	X
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-

7

Figures

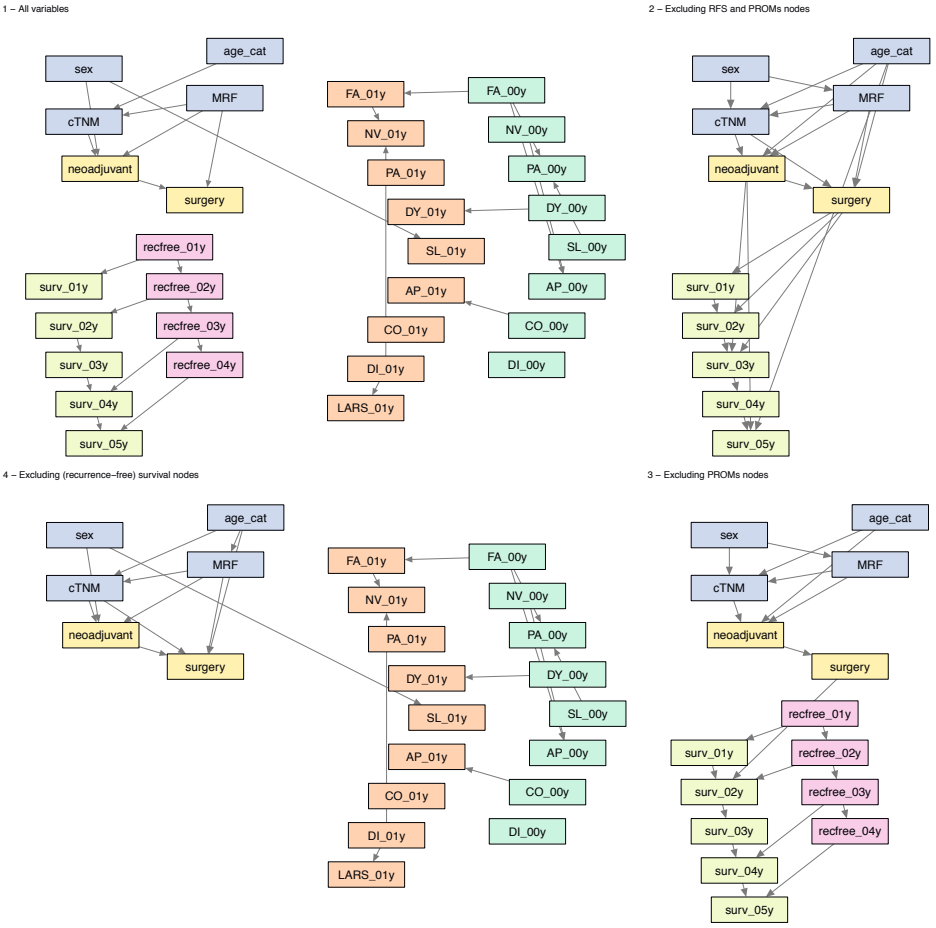


Figure 1: *Top left:* Initial Bayesian Network, after structure learning using blacklist only. Treatment (yellow nodes) are not associated with recurrence-free survival (purple nodes) or overall survival (light-green nodes). There is no association between PROMs nodes (green and orange nodes) and treatment, recurrence or survival. MRF: mesorectal fascia; FA: Fatigue; NV: Nausea/Vomiting; PA: Pain; DY: Dyspnea; SL: Sleep Disturbances; AP: Appetite Loss; CO: constipation; DI: Diarrhea; LARS: Low Anterior Resection Syndrome score. The suffix _0{n}y for recurrence, survival, and PROMs, indicates the number of years after diagnosis. *Top right:* Bayesian Network after structure learning using a blacklist and excluding recurrence-free survival and PROMs nodes. *Bottom right:* Bayesian Network after structure learning using a blacklist and excluding PROMs nodes. *Bottom left:* Bayesian Network after structure learning using a blacklist and excluding recurrence-free survival and overall survival nodes.

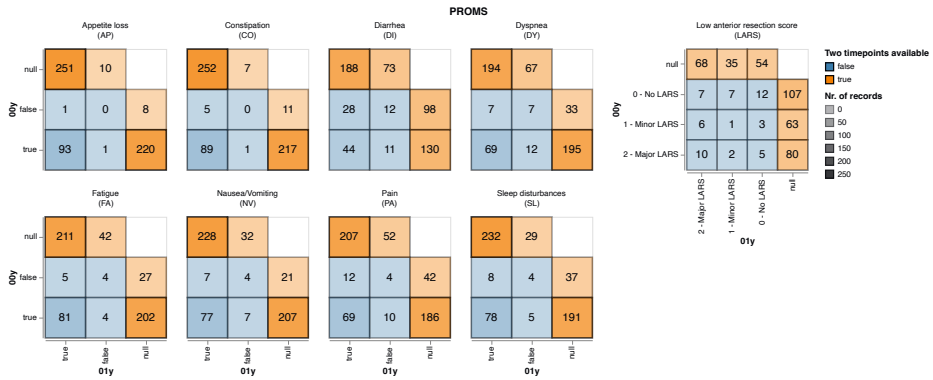


Figure 2: Distribution of PROMs domains, dichotomized as "true" (clinically important) or "false" (clinically unimportant), at baseline and 1 year after diagnosis. A blue background color indicates two timepoints were available for a single patient; orange indicates either a baseline measurement or a measurement 1 year after diagnosis was available. This shows that for most records only a single measurement is available.

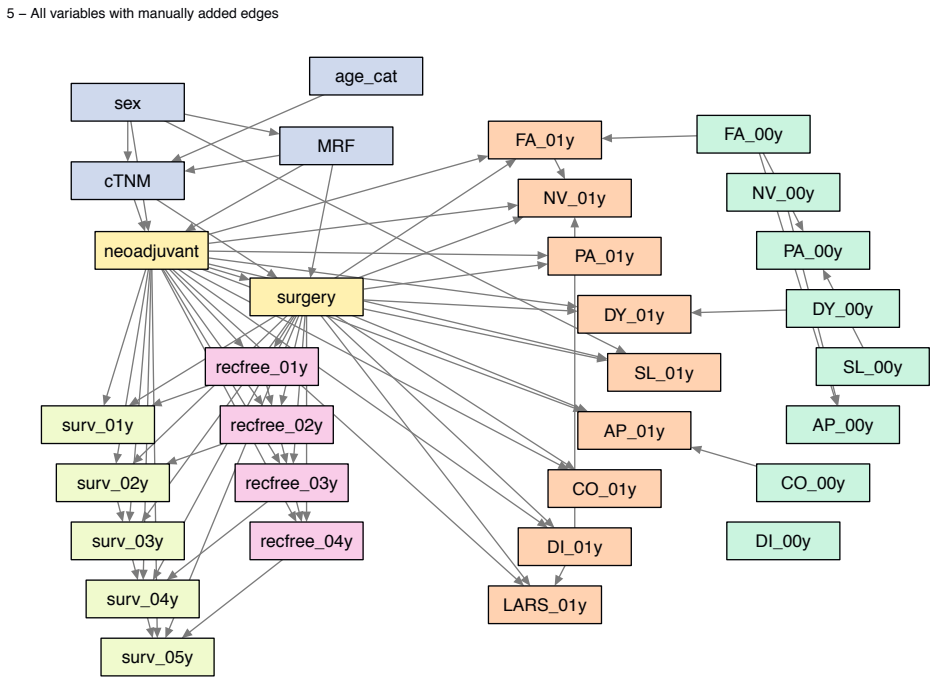


Figure 3: Final Bayesian Network structure after structure learning using a blacklist and manually adding edges between treatment nodes (neoadjuvant and surgery, sex → cTNM, and sex → MRF. Yellow nodes: treatment. Purple nodes: recurrence-free survival. Light green nodes: overall survival. Green nodes: baseline PROMs. Orange nodes: PROMs 1-year after diagnosis. MRF: mesorectal fascia; FA: Fatigue; NV: Nausea/Vomiting; PA: Pain; DY: Dyspnea; SL: Sleep Disturbances; AP: Appetite Loss; CO: constipation; DI: Diarrhea; LARS: Low Anterior Resection Syndrome score. The suffix _0{n}y for recurrence, survival, and PROMs, indicates the number of years after diagnosis.

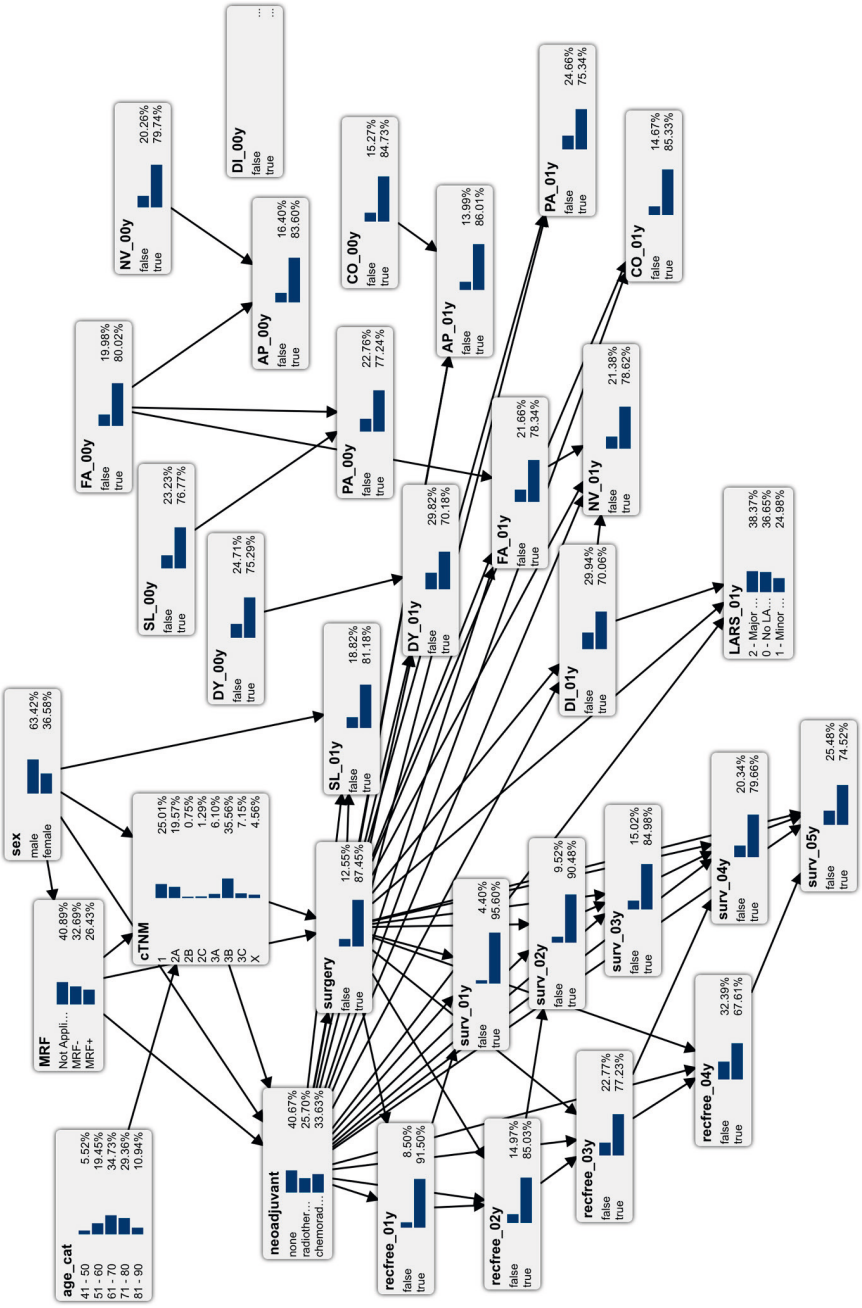


Figure 4: Final Bayesian Network, additionally showing prior probabilities for each node, after structure learning using a blacklist and manually adding edges between treatment nodes (neoadjuvant and surgery, sex → cTNM, and sex → MRF. MRF: mesorectal fascia; cTNM: clinical TNM-stage group, recfree: recurrence-free survival; surv: survival. FA: Fatigue; NV: Nausea/Vomiting; PA: Pain, DY: Dyspnea, SL: Sleep Disturbances; AP: Appetite Loss; CO: constipation; DI: Diarrhea; LARS: Low Anterior Resection Syndrome score. The suffix _0{n}y for recurrence, survival, and PROMs, indicates the number of years after diagnosis.

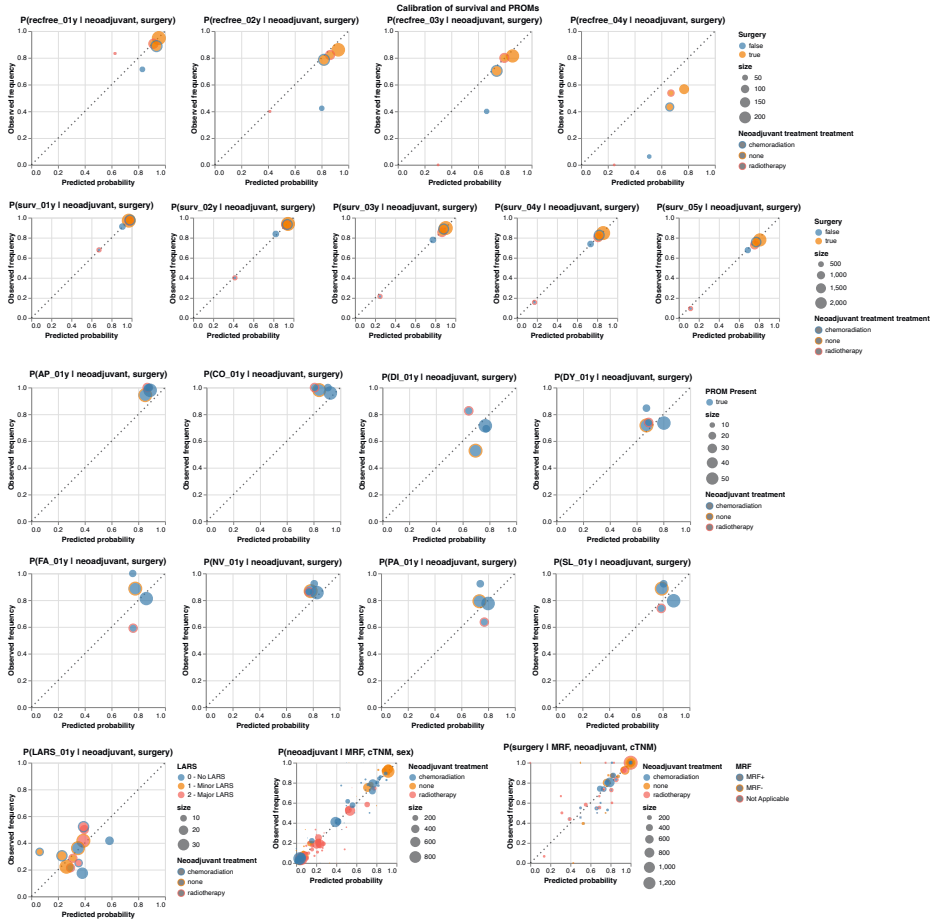


Figure 5: Visualization of model calibration for recurrence-free survival, survival, PROMs, neoadjuvant treatment, and surgery. Each panel corresponds to a node in the Bayesian Network. Within a panel, each circle represents a specific subpopulation as defined by the states of (a subset of) the node's parents and the states of the node itself. If a node only had two states (Boolean), only the "true" stage is shown (e.g. for 1-year survival only the subgroup where "true" is predicted is shown). For each subgroup, predicted probability (x-axis) is plotted against the observed frequency (y-axis) in the dataset. In case of the (recurrence-free) survival and PROMs domain, the parents neoadjuvant and surgery were selected. For nodes neoadjuvant and surgery their respective parents were selected.

References

1. NKR Incidentie Rectumcarcinoom 2007-2017 [Internet]Available from: https://www.iknl.nl/nkr-cijfers?fs%7Cepidemiologie_id=6&fs%7Ctumor_id=217&fs%7Cregio_id=155&fs%7Cperiode_id=96%2C97%2C98%2C99%2C100%2C101%2C102%2C103%2C104%2C105%2C106&fs%7Cgeslacht_id=15&fs%7Cleeftijdsgroep_id=67&fs%7Cjaren_na_diagnose_id=16&fs%7Ceenheid_id=2&cs%7Ctype=column&cs%7CxAxis=periode_id&cs%7Cseries=epidemiologie_id&ts%7CrowDimensions=periode_id&ts%7CcolumnDimensions=&lang%7Clanguage=nl
2. van Gijn W, Marijnen CAM, Nagtegaal ID, et al: Preoperative radiotherapy combined with total mesorectal excision for resectable rectal cancer: 12-year follow-up of the multicentre, randomised controlled TME trial. Lancet Oncol 12:575–582, 2011
3. Sauer R, Becker H, Hohenberger W, et al: Preoperative versus postoperative chemoradiotherapy for rectal cancer. N Engl J Med 351:1731–1740, 2004
4. Sauer R, Liersch T, Merkel S, et al: Preoperative versus postoperative chemoradiotherapy for locally advanced rectal cancer: results of the German CAO/ARO/AIO-94 randomized phase III trial after a median follow-up of 11 years. J Clin Oncol Off J Am Soc Clin Oncol 30:1926–1933, 2012
5. Schlechter BL: Management of Rectal Cancer. Hematol Oncol Clin North Am 36:521–537, 2022
6. Marijnen C a. M, Kapiteijn E, van de Velde CJH, et al: Acute side effects and complications after short-term preoperative radiotherapy combined with total mesorectal excision in primary rectal cancer: report of a multicenter randomized trial. J Clin Oncol Off J Am Soc Clin Oncol 20:817–825, 2002
7. Ankolekar A, Dekker A, Fijten R, et al: The Benefits and Challenges of Using Patient Decision Aids to Support Shared Decision Making in Health Care. JCO Clin Cancer Inform 1–10, 2018
8. Ogura A, Shiomi A, Yamamoto S, et al: Prediction model of the risk for lateral local recurrence in locally advanced rectal cancer without enlarged lateral lymph nodes: Lessons from a Japanese multicenter pooled analysis of 812 patients. Ann Gastroenterol Surg 8:284–292, 2024
9. Jeon Y, Kim Y-J, Jeon J, et al: Machine learning based prediction of recurrence after curative resection for rectal cancer. PloS One 18:e0290141, 2023
10. Hida K, Okamura R, Park SY, et al: A New Prediction Model for Local Recurrence After Curative Rectal Cancer Surgery: Development and Validation as an Asian Collaborative Study. Dis Colon Rectum 60:1168–1174, 2017
11. Orive M, Anton-Ladislao A, Lázaro S, et al: Anxiety, depression, health-related quality of life, and mortality among colorectal patients: 5-year follow-up. Support Care Cancer Off J Multinatl Assoc Support Care Cancer 30:7943–7954, 2022
12. Choy I, Young JM, Badgery-Parker T, et al: Baseline quality of life predicts pelvic exenteration outcome. ANZ J Surg 87:935–939, 2017
13. Liu H, Lv L, Qu Y, et al: Prediction of cancer-specific survival and overall survival in middle-aged and older patients with rectal adenocarcinoma using a nomogram model. Transl Oncol 14:100938, 2021
14. Aaronson NK, 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. JNCI J Natl Cancer Inst 85:365–376, 1993
15. Giesinger JM, Loth FLC, Aaronson NK, et al: Thresholds for clinical importance were established to improve interpretation of the EORTC QLQ-C30 in clinical practice and research. J Clin Epidemiol 118:1–8, 2020
16. Sieswerda M, Xie S, van Rossum R, et al: Identifying Confounders Using Bayesian Networks and Estimating Treatment Effect in Prostate Cancer With Observational Data. JCO Clin Cancer Inform 7:e2200080, 2023
17. Sieswerda M, Rossum R van, Bermejo I, et al: Estimating treatment effect of adjuvant chemotherapy in elderly stage III colon cancer patients using Bayesian Networks. JCO Clin Cancer Inform , 2023

18. Dempster AP, Laird NM, Rubin DB: Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc Ser B Methodol 39:1–38, 1977
19. Sprangers MA, Schwartz CE: The challenge of response shift for quality-of-life-based clinical oncology research. Ann Oncol Off J Eur Soc Med Oncol 10:747–749, 1999
20. Sawatzky R, Sajobi TT, Russell L, et al: Response shift results of quantitative research using patient-reported outcome measures: a descriptive systematic review. Qual Life Res Int J Qual Life Asp Treat Care Rehabil 33:293–315, 2024
21. Cijfers darmkanker [Internet][cited 2024 Jul 1] Available from: https://iknl.nl/kankersoorten/darmkanker/registratie
22. Heald RJ, Ryall RDH: RECURRENCE AND SURVIVAL AFTER TOTAL MESORECTAL EXCISION FOR RECTAL CANCER. The Lancet 327:1479–1482, 1986

Supplemental Materials

Tables

variable	values	neoadj. chemoradiation (n=4632)	neoadj. radiotherapy (n=3382)	surgery only (n=5775)	p	sign.
age	**41 - 50**	437 (9.4)	117 (3.5)	183 (3.2)	0.0000	***
	51 - 60	1222 (26.4)	552 (16.3)	937 (16.2)		
	61 - 70	1600 (34.5)	971 (28.7)	2188 (37.9)		
	71 - 80	1198 (25.9)	971 (28.7)	1869 (32.4)		
	81 - 90	175 (3.8)	771 (22.8)	598 (10.4)		
sex	**male**	2887 (62.3)	2138 (63.2)	3741 (64.8)	0.0315	*
	female	1745 (37.7)	1244 (36.8)	2034 (35.2)		
cTNM	**1**	126 (2.7)	160 (4.7)	3237 (56.1)	0.0000	***
	2A	601 (13.0)	604 (17.9)	1490 (25.8)		
	2B	58 (1.3)	28 (0.8)	12 (0.2)		
	2C	114 (2.5)	54 (1.6)	15 (0.3)		
	3A	111 (2.4)	562 (16.6)	133 (2.3)		
	3B	2784 (60.1)	1787 (52.8)	261 (4.5)		
	3C	826 (17.8)	129 (3.8)	24 (0.4)		
	X	12 (0.3)	58 (1.7)	603 (10.4)		
MRF	**Not Applicable**	861 (19.9)	893 (28.5)	3892 (70.3)	0.0000	***
	MRF-	1093 (25.3)	1563 (49.9)	1388 (25.1)		
	MRF+	2362 (54.7)	678 (21.6)	253 (4.6)		
recfree_01y	**FALSE**	55 (10.6)	37 (11.5)	29 (5.1)	0.0005	**
	TRUE	464 (89.4)	286 (88.5)	544 (94.9)		
recfree_02y	**FALSE**	124 (24.5)	66 (21.6)	69 (12.7)	0.0000	***
	TRUE	383 (75.5)	240 (78.4)	473 (87.3)		
recfree_03y	**FALSE**	158 (31.9)	78 (26.5)	92 (18.1)	0.0000	***
	TRUE	337 (68.1)	216 (73.5)	416 (81.9)		
recfree_04y	**FALSE**	164 (62.1)	87 (52.7)	104 (40.8)	0.0000	***
	TRUE	100 (37.9)	78 (47.3)	151 (59.2)		
recfree_05y	**FALSE**	168 (90.8)	90 (84.9)	110 (77.5)	0.0037	**
	TRUE	17 (9.2)	16 (15.1)	32 (22.5)		
surv_01y	**FALSE**	167 (3.6)	278 (8.2)	167 (2.9)	0.0000	***
	TRUE	4463 (96.4)	3103 (91.8)	5607 (97.1)		
surv_02y	**FALSE**	411 (8.9)	567 (16.8)	335 (5.8)	0.0000	***
	TRUE	4216 (91.1)	2812 (83.2)	5434 (94.2)		
surv_03y	**FALSE**	650 (14.1)	854 (25.3)	571 (9.9)	0.0000	***

variable	values	neoadj. chemoradiation (n=4632)	neoadj. radiotherapy (n=3382)	surgery only (n=5775)	p	sign.
	TRUE	3975 (85.9)	2525 (74.7)	5195 (90.1)		
surv_04y	FALSE	911 (19.7)	1047 (31.0)	841 (14.6)	0.0000	***
	TRUE	3714 (80.3)	2331 (69.0)	4922 (85.4)		
surv_05y	FALSE	1096 (26.7)	1208 (38.9)	1091 (21.4)	0.0000	***
	TRUE	3013 (73.3)	1900 (61.1)	4005 (78.6)		
FA_00y	FALSE	21 (15.9)	6 (8.0)	9 (7.8)	0.0774	
	TRUE	111 (84.1)	69 (92.0)	107 (92.2)		
FA_01y	FALSE	21 (13.0)	16 (25.0)	13 (10.7)	0.0246	*
	TRUE	141 (87.0)	48 (75.0)	108 (89.3)		
NV_00y	FALSE	14 (10.6)	6 (8.0)	12 (10.3)	0.8176	
	TRUE	118 (89.4)	69 (92.0)	104 (89.7)		
NV_01y	FALSE	21 (12.7)	9 (13.6)	13 (10.6)	0.7928	
	TRUE	145 (87.3)	57 (86.4)	110 (89.4)		
PA_00y	FALSE	34 (25.8)	8 (10.7)	16 (13.8)	0.0086	*
	TRUE	98 (74.2)	67 (89.3)	100 (86.2)		
PA_01y	FALSE	30 (18.2)	14 (21.2)	22 (17.9)	0.8366	
	TRUE	135 (81.8)	52 (78.8)	101 (82.1)		
DY_00y	FALSE	16 (12.1)	10 (13.3)	21 (18.1)	0.3880	
	TRUE	116 (87.9)	65 (86.7)	95 (81.9)		
DY_01y	FALSE	35 (21.1)	18 (26.9)	33 (26.8)	0.4486	
	TRUE	131 (78.9)	49 (73.1)	90 (73.2)		
SL_00y	FALSE	24 (18.2)	10 (13.3)	15 (12.9)	0.4541	
	TRUE	108 (81.8)	65 (86.7)	101 (87.1)		
SL_01y	FALSE	18 (10.8)	10 (14.9)	10 (8.1)	0.3482	
	TRUE	148 (89.2)	57 (85.1)	113 (91.9)		
AP_00y	FALSE	6 (4.5)	0 (0.0)	3 (2.6)	0.1592	
	TRUE	126 (95.5)	75 (100.0)	113 (97.4)		
AP_01y	FALSE	8 (4.8)	0 (0.0)	3 (2.4)	0.1375	
	TRUE	158 (95.2)	67 (100.0)	120 (97.6)		
CO_00y	FALSE	9 (6.8)	5 (6.7)	2 (1.7)	0.1345	
	TRUE	123 (93.2)	70 (93.3)	114 (98.3)		
CO_01y	FALSE	4 (2.4)	3 (4.5)	1 (0.8)	0.2532	
	TRUE	161 (97.6)	63 (95.5)	122 (99.2)		
DI_00y	FALSE	65 (49.2)	33 (44.0)	40 (34.5)	0.0620	
	TRUE	67 (50.8)	42 (56.0)	76 (65.5)		
DI_01y	FALSE	38 (22.9)	18 (26.9)	40 (32.5)	0.1895	

variable	values	neoadj. chemoradiation (n=4632)	neoadj. radiotherapy (n=3382)	surgery only (n=5775)	p	sign.
	TRUE	128 (77.1)	49 (73.1)	83 (67.5)		
LARS_00y	**0 - No LARS**	36 (31.0)	36 (49.3)	61 (53.5)	0.0004	***
	1 - Minor LARS	26 (22.4)	21 (28.8)	26 (22.8)		
	2 - Major LARS	54 (46.6)	16 (21.9)	27 (23.7)		
LARS_01y	**0 - No LARS**	33 (39.3)	12 (26.7)	29 (35.8)	0.5529	
	1 - Minor LARS	15 (17.9)	13 (28.9)	17 (21.0)		
	2 - Major LARS	36 (42.9)	20 (44.4)	35 (43.2)		

Table 1: Full variable distribution and association with treatment.

Figures

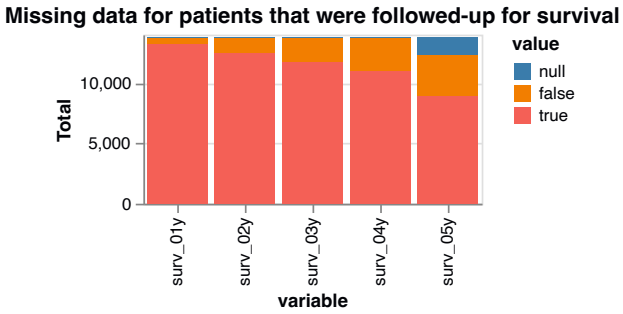


Figure S1: Distribution of missing data in the (discretized) survival columns

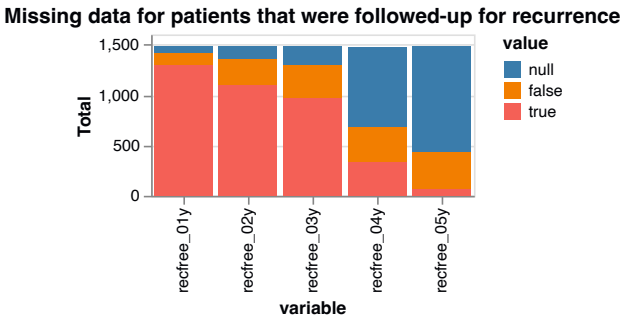


Figure S2: Distribution of missing data in the (discretized) recurrence columns for patients that were followed up for recurrence only.

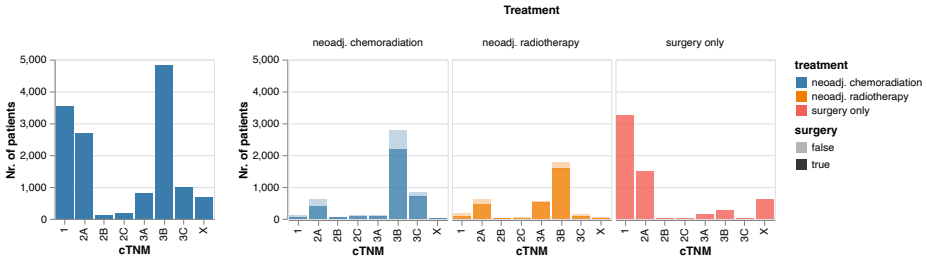


Figure S3: *Left:* Distribution of clincial tumor stage (cTNM). *Right:* Distribution of clinical tumor stage, stratified by initial (neoadjuvant) treatment and (subsequent) surgery.

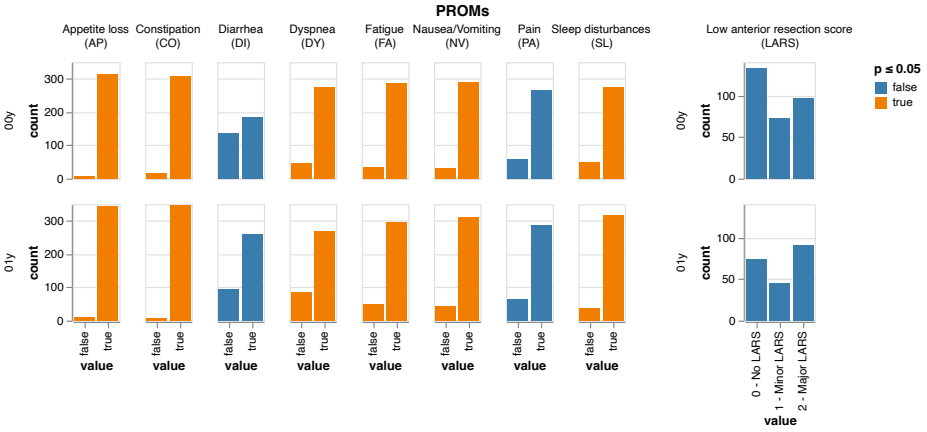


Figure S4: Histograms of dichotomized PROMs domains at baseline (top row) and 1 year after diagnosis. Color indicates statistically significant association between baseline and effect (at $p \leq 0.05$) as determined using the Chi^2 test.

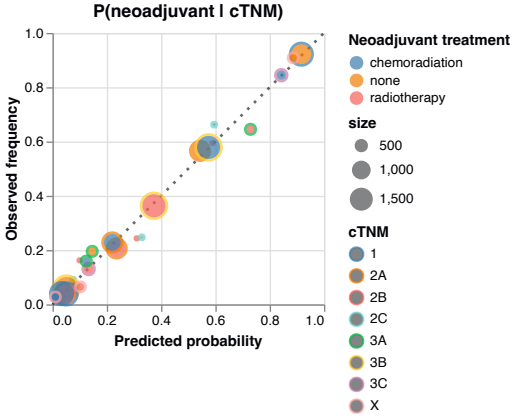


Figure S5: Calibration of neoadjuvant treatment given cTNM stage.

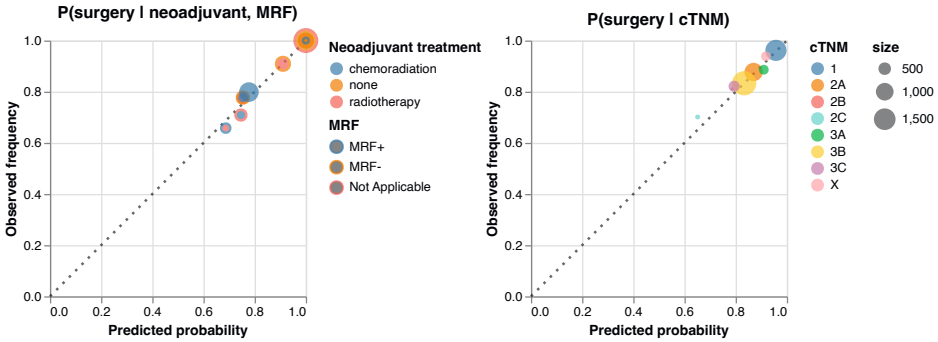


Figure S6: Calibration of surgery given either neoadjuvant and MRF, or cTNM

8 Discussion

Introduction

Real world data are key to improving healthcare. They can be used to monitor disease incidence and prevalence, organization and quality of care, and trends in treatment. Additional uses include exploratory research and improving patient outcomes. With the advent of Electronic Healthcare Records (EHRs), and adoption of computers in general, both the interest in and the possibility for data reuse have increased.

Extracting routine clinical data for re-use in research

For every use-case that involves data reuse, (some form of) standardization is a prerequisite. Specifically, in Chapter 2 we have shown that it is possible to adapt EHRs and that standardized, structured tumor board reporting can be implemented in complex, multidisciplinary processes. In this case, the adaptation consisted of redesigning and implementing several tumor board forms to align with the national information standard for breast cancer that, in turn, was derived from the (electronic version of the) Dutch Clinical Practice Guideline (CPG). The implementation benefited the quality of tumor board reporting, enabled data reuse for research, and did not result in additional clinical workload.

While the project was successful, it hinged on the availability of a national information standard for breast cancer, derived from (computer interpretable) CPGs.[1] Development of this information standard alone, from initial idea to nationally approved complement to the guideline, took significant effort and was made possible by the fact that breast cancer care in the Netherlands follows well-defined processes. Not all clinical processes are so well-defined, and CPGs may not be available. Furthermore, decision points are not always so explicitly marked. This results in limited generalizability and data reusability.

Historically, structuring and standardization of clinical data have taken place to solve problems in the primary care process. For example, using discrete fields enabled graphical representation of a patient's status (e.g., line charts of vital signs) which yielded a clinical benefit. Financial reasons were another driver: standardization of diagnoses and procedures facilitated processing, reimbursement by healthcare insurance companies, and billing. Finally, interoperability between (ancillary) systems within the hospital required standardization, but provided the ability to view lab results electronically in tabular form as soon as they became available.[2] Standardization was carried out to the point of (clinical) relevance and/or benefit. For example, narrative data like clinical notes or radiology reports, remained largely in free text. As a result, EHR data is a mix of structured and unstructured data. In the Netherlands, attempts to further standardize clinical content, especially narrative data, have met with limited success.[3] This has several reasons.

First, standardization is costly and time consuming, proportional to the number of participants involved in the process. As a result, while diagnoses and procedures have been standardized at a national level because it was required for reimbursement, local standardization was sufficient to support communication between ancillary systems. Where agreement can be found on standardization, implementation on a national level may require support from EHR vendors which can present another obstacle. Of course, local standards are better than no standards, but in the absence of (inter)national standards it remains difficult to reuse or exchange data between different sites.

Furthermore, requirements between primary and secondary use are difficult to align: clinicians prefer free text for recording narrative data as it provides the flexibility needed, but secondary use (e.g., research, billing, and quality and process monitoring) typically requires well defined variables (which are usually a derivative of what is needed clinically).[4, 5]

Finally, further standardization provides no (or little) clinical benefit: free text is fine when intended for human consumption. Standardization may even bring additional workload, as applying classification can be time consuming if not properly aligned with the clinical process. From a secondary use perspective, this is unfortunate. The narratives are where most of the clinical information comes together and are where context, conclusions, and decisions are captured.

Given these difficulties, it is interesting to reflect on why standardizing tumor board reporting was successful at Amphia hospital. In our opinion, several factors can be identified. Firstly, the clinical documentation was part of a well-defined process, outlined by the national clinical practice guideline (CPG) which described both workup and decision points. Secondly, the variables required for decision making were derived from the guidelines and therefore, by definition, aligned with the clinical process. Finally, the auto-generated, textual note provided clinical benefit as it simplified communication with colleagues, such as general practitioners.

All these factors ensured that standardized structured reporting did not put any additional burden on clinicians, as shown by required preparation time. In fact, the project yielded clinical benefit through improvement of clinical documentation and communication. The above is in line with recent efforts by Ebbers et al. where a care pathway was used with even more pronounced results.[6]

While our approach shows promise for the future (under the right circumstances it *is* possible to fulfil both primary and secondary use requirements) another route has recently materialized. The latest Large language models (LLMs), such as ChatGPT and GPT-4, can be leveraged to interpret free text and extract relevant information from clinical notes, radiology reports, and other sources of information with a high degree of

accuracy.[7, 8] For example, Fink et al. reported that GPT-4 was able to extract the diameter of primary tumors with 98.8% accuracy from free-text CT-reports.[9] Although application of LLMs in this fashion does not come with the benefit of better documentation, it does appear to be a quicker and more scalable solution for obtaining discrete data for research compared to our approach.

Dealing with changes in classification systems (over time)

Next, in Chapter 3 we illustrated how Bayesian networks can be used to reclassify stage groupings across TNM-editions by combining clinical knowledge with real-world data.

This is important because one of the factors slowing down research is the time required from intervention to outcome measure: determining 5-year overall survival usually requires 5-year follow-up. Of course, surrogate endpoints, such as recurrence-free survival, typically require shorter follow-up time and are frequently used to hasten research.[10, 11] However, recurrence-free survival does not, unfortunately, always correlate with overall survival.[12]

Routinely collected observational data provide a different way to circumvent the need to wait for follow-up by simply looking at the past. In oncology, an important variable (or set of variables) that is routinely documented is the TNM-classification. As explained in the Introduction and Chapter 3, the TNM-classification system guides treatment decisions, aids in stratifying patients for research, and helps clinicians assess prognosis.[13, 14] To keep the system current, it is updated on a regular basis, but updates are not always backwards compatible.

Since TNM-stage is associated with both treatment selection and outcome, it plays a pivotal role in research and acts as a potential confounder (see also Introduction, Chapter 5, Chapter 6, and Chapter 7). If the variable representing tumor stage group consists of a mixture of editions, this may result in difficulty in detecting the true treatment.

Possible solutions to this problem include mapping data from one classification system version to the other, basing the reclassification on underlying data, and discarding part of the data. Unfortunately, mapping is not always possible, underlying data is not always available, and while discarding part of the data might provide a solution from a statistical point of view, it reduces the available data which introduces limitations of its own.

Our approach, to level and equalize the data across TNM-editions through probabilistic reclassification, leveraged both knowledge about changes between the editions of the classification system and the correlation between the TNM-stage and survival. This provides an alternative way to increase available data and helps data retain its value.

While our method was successful, evaluation of the resulting models remains a fundamental issue (essentially a Catch-22): creation of a test set runs into the same issues that required development of our method in the first place. And in the absence of a test set, uncertainty about the usefulness of the model remains.

In the specific situation described in Chapter 3 we were able to create a test set by reclassifying part of the data using rules, since some of the underlying variables were available. Reclassification of individual TNM-descriptors was not sufficiently reliable, but TNM-stage grouping *could* be tested. If this had not been possible, an alternative solution might have been to go back to the data source (the EHR) and attempt to obtain the data required for reclassification of the test set. Even though this would have come with significant effort, it would have still been more economical than reclassifying the entire dataset in this manner. Of course, in an ideal world, classification systems would be backward compatible. Whether this is feasible, depends on the classification system in question and the pressures it is facing to change.

Using (clinical) data across borders with Federated Learning

In the subsequent chapter, Chapter 4, we have shown how Federated Learning can be used to learn from data across borders while simultaneously adhering to laws and regulations, and mitigating privacy concerns. Not only can this technique be applied to increase available data, but it also unlocks comparisons between widely different practices. For example, in this specific study, cultural differences appear to have a strong impact on the incidence of oral cavity cancer and (as a result) on clinical practice guidelines.

To perform the analyses, we implemented a federated version of the Cox proportional hazards algorithm as described by Lu et al.[15] Unfortunately, the algorithm to check the proportional hazards assumption was not yet implemented, and the assumption could not be checked. This can be considered a limitation of our study.

Interpretation of the results presented another challenge.

Identifying plausible (causal) explanations can be considered a difficult task under any circumstance, but here the uncertainty was more strongly felt because potential explanations were not only rooted in (patho)physiology, but included variables related to process and experience.

At the same time, because of the widely different practices between the countries, there was no intuition with respect to causes and probabilities. Here, federated algorithms for Bayesian Networks might have helped in identifying and mitigating confounders, and to get a better understanding of the underlying causes.[16, 17]

Technical barriers that existed previously, for example the availability of software to setup an infrastructure and accompanying algorithms, have been largely mitigated since our study took place.[18–20] The technique has matured and can be considered ready for production use.

However, despite these efforts, some challenges remain. Specifically, although in Federated Learning data is not shared in the traditional sense, collaborations still need to agree to use each other's data and any applicable limitations. This requires a certain level of understanding of the technique, its pitfalls, and (legal) risks. Understandably, this is (still) new terrain for legal departments, which tends to prolong turnaround time. Additionally, data standardization and metadata are even more important in Federated Learning than in centralized analyses since meaning of the data cannot be inferred by simply looking at the data. Alignment of data within a collaboration can still take significant effort. Nevertheless, once these hurdles have been taken, Federated Learning facilitates privacy by design, access to data previously unavailable, and eliminates the need for data transport.

Internationally, the number of initiatives that aim to develop a sustainable federated infrastructure are rapidly growing, though usually targeting a very specific domain.[21–23] A national initiative to make healthcare data available for federated analysis, along with the required rules, regulations, and standard contracts, would both simplify and quicken research.

Identifying and mitigating confounders to estimate treatment effect

When it comes to using observational data for comparative effectiveness research, there are two major hurdles: identifying confounding variables in the dataset and the possibility of unmeasured confounding.

In 2008, the importance of overcoming these hurdles was illustrated by Giordano et al.[24] They showed, by analyzing data from the Surveillance, Epidemiology, and End Results (SEER) Program (US), that failing to address these hurdles leads to implausible results. For example, they found androgen deprivation was associated with a higher risk of death in locally advanced (stage III) prostate cancer which is contrary to well-established clinical evidence. They also determined that, in localized prostate cancer, active treatment was associated with a lower hazard for *other-cause* mortality and that patients that underwent radical prostatectomy had a better prognosis than a control population *without* cancer. Regular methods for dealing with confounding, such as propensity score correction, proved insufficient in their study.[25, 26]

Therefore, in Chapters 5 and 6, we investigated how observational data can be used to identify and correct for confounders, ultimately obtaining causal models. These causal models allow for comparison of treatment effect, which reduces the need for randomized controlled trials (RCTs) and greatly enhances the value of the data. Especially since RCTs are not always feasible or ethical, and typically only cover a small part of the population.

The method we developed to obtain these causal models makes use of structure learning algorithms for Bayesian networks and is conceptually simple. By precluding directed edges that cannot possibly exist (e.g., tumor characteristics never *cause* age) one can assume that any edges found by structure learning are likely to be causal. The resulting causal model can subsequently be used to identify confounders, after which mitigation is relatively straightforward. In this way, we combined (basic) clinical knowledge with real world data.

In Chapter 5, our first application of the above method, we compared one of the analyses Giordano et al. performed using SEER data with a causal analysis of data from the Netherlands Cancer Registry (NCR): we investigated the effect of active treatment compared to observation in localized prostate cancer. This form of prostate cancer can run an indolent clinical course where the benefit of active treatment, such as radical prostatectomy or radiotherapy, is limited.[27]

Interestingly, our method did not identify socioeconomic status (SES) as an independent cause of either treatment or survival. Additionally, the resulting Bayesian Network suggested that, while PSA (Prostate Specific Antigen) appeared to be a driver for treatment selection, PSA did not act as an independent cause of survival. While these variables are often considered confounders, this result suggests correcting for them is unnecessary.

We found that active treatment had a significant, but limited effect in our cohort: a 1 - 3% increase in 10-year overall survival. This is in line with what is reported in literature, but lower than previous estimates based on data from the SEER Program.[24, 28, 29] The disparity might be explained by differences in screening practice and healthcare systems between the US and the Netherlands. We concluded that we were able to successfully correct for confounding in our dataset.

In Chapter 6 we investigated the effect of adjuvant chemotherapy in elderly stage III colon cancer patients. Here, Dutch guidelines recommend treatment with surgery and adjuvant chemotherapy, but in elderly patients, the benefit of regimens containing oxaliplatin is still a point of contention.[30, 31] Previously, we found that treatment with either capecitabine and oxaliplatin (CAPOX) or capecitabine monotherapy (CapMono) was associated with improved recurrence-free survival (RFS) and overall survival (RFS); superiority for either regimen could not be established. At the time, we adjusted for confounding by indication using the pretreatment criterion but did not apply causal analysis.

In this more recent study, we revisited the data using the method described above, and estimated treatment effect of adjuvant chemotherapy after identifying potential confounders. Additionally, we explored the effect of using three different structure learning algorithms on confounder identification.

Indeed, different structure learning algorithms identified different potential (sets of) confounders, but none were contradicting. In other words, confounders identified in one model were not mediators or colliders in others.

Apart from the aforementioned confounders, the Bayesian Networks found several interesting associations between the (pretreatment) variables. For example, sex was associated with tumor location and grade with pN; associations that have both been reported before.[32, 33] This further emphasizes how algorithms like ours can be used to elucidate underlying relationships. With respect to treatment effect, we found a strong association between capecitabine and survival in our cohort, but no additional benefit of oxaliplatin was found.

The results described in Chapters 5 and 6 show that the restraint that is typically employed when it comes to observational data (i.e., correlation is not causation) is not always warranted. The ability to compare treatment effect using observational data greatly enhances its use. This conclusion is supported by a Cochrane review from 2014, that found that, "on average, there is little difference between the results obtained from RCTs and observational studies".[34]

Still, although unmeasured confounding did not appear to be an issue in our analyses, the uncertainty that it may be present remains a challenge for subsequent analyses on different data. For example, our method was applied to oncological data from the Netherlands where oncological care is well organized, and CPGs drive many decisions. In the Netherlands, health insurance is mandatory and oncological care is fully covered (without significant co-pay). While socioeconomic differences certainly exist and may determine prior health status, they generally do not limit access to diagnostics and treatment options. It can be reasoned that this reduces the amount of (unmeasured) confounding in our data compared to countries with larger inequalities.

However, when specifically considering treatment selection, one could argue that clinicians are likely to be aware of potential confounders as they discuss treatment options with their patients. In that respect, assessing availability of all relevant variables is not unlike clinical reasoning. Our method of confounder identification has the advantage that clinical input is 'built in'.

Simultaneously, this dependency on clinical (prior) knowledge can be considered a limitation: the availability of prior (clinical) knowledge is strongly linked to the nature of

the data. If, for instance, our method would be applied to genetics data, both the number of variables and identifying forbidden edges would present significant challenges.

Towards clinical application

While machine learning models are useful in a research setting and for guiding clinical practice (e.g., as input during CPG development), they are increasingly being investigated for use at the point of care as Clinical Decision Support (CDS) systems.[35] However, even though the availability of CDS systems increases, adoption into routine clinical care remains behind.[36–38] One of the underlying reasons is that many CDS systems/models are essentially black boxes to the end user. Not being able to understand what drives a model's alerts and suggestions, does not inspire trust.[35]

Bayesian Networks (BNs) might mitigate some of these barriers. For example, they have the advantage of being graphical models, which allows visual representation of the model. This includes both relationships between variables and visualization of probability distributions, which makes it straightforward for physicians to check the model's assumptions and compare them with their own understanding of the medical domain. Furthermore, their graphical representation ties in with their ability to support causal reasoning while simultaneously giving users an intuitive understanding of what goes on in the background.

Additionally, BNs are capable of dealing with missing values and support real-time evidence propagation; both these features are useful in a clinical context where understanding of a patient's illness is frequently a dynamic process.

In Chapter 7 we developed a Bayesian Network to help in shared decision making in rectal cancer, which is the 5th-most frequently occurring malignancy in the Netherlands with an annual incidence around 4700 new cases a year. In stage II – III rectal cancer, Dutch clinical practice guidelines recommend neoadjuvant (chemo)radiotherapy. Ultimately, however, the choice for (a specific type of) neoadjuvant therapy is a shared decision between doctor and patient. Through shared decision making (SDM), a balance is sought between the risks and benefits for a particular treatment. For neoadjuvant therapy in rectal cancer specifically, these choices revolve around local control (preventing local recurrence), patient reported outcome measures (PROMS) and Quality of Life (QoL).

As a first step towards supporting effective SDM surrounding neoadjuvant (chemo) radiotherapy in patients with rectal cancer, we set out to develop a model that simultaneously predicts recurrence-free survival, overall survival, and quality of life. In our study, however, initially we were not able to identify associations between treatment choice and recurrence, survival, or quality of life. Association between treatment and survival only appeared after we left the recurrence-free survival out of the model. We

believe this can be explained by the limited amount of PROMS- and recurrence data that was available, in conjunction with a relatively short follow-up.

By manually adding edges after structure learning, we still obtained a model that could simultaneously predict recurrence, survival, and quality of life. Calibration plots showed that the model prediction for 2- and 3-year survival was well calibrated, but prediction for 4- and 5-year survival deteriorated, likely as a result of the aforementioned missing values. Calibration of PROMS-prediction differed per domain. However, as could be expected since the BN did not find any edges between treatment and PROMS, in general the predictions were fairly close together. This was especially the case for the domains of appetite loss and constipation.

Unfortunately, BNs have downsides too. Currently the implementations for querying BNs in common statistical packages, such as Hugin, bnlearn, or pgmpy, do not offer confidence intervals for their predictions.[39–41] Having these available, would give clinicians a better intuition about the relevance of a prediction. This becomes especially important in either edge cases, or models with many interlinked variables and thus high dimensionality, where insufficient (training) data is available for parameter estimation.

In the (near) future, both standardization and techniques such as Federated Learning might mitigate this issue of data availability.

Conclusion

In the previous chapters we have seen how standardization of electronic healthcare records may be achieved without negatively impacting clinical workflow data and subsequently how can be extracted from the clinical process by leveraging standardized tumor board reporting. Next, we have shown how we can ensure that older data retains its value. Furthermore, we have shown how patient privacy concerns can be mitigated and how data may be (safely) used across international borders by using Federated Learning. Finally, and perhaps most importantly, we have illustrated how we can leverage observational data to reliably estimate treatment effect with causal inference, thus greatly enhancing its value. Generalizability of this method, however, remains uncertain as it depends on domain expertise and, ultimately, the *possibility* of unmeasured confounding can never be fully ruled out. Here, further research is required to establish validity of the method under other circumstances (or for other tumors) and generate trust.

While we had hoped to apply our method to develop a model that could be used for Shared Decision Making, this proved a bridge too far due to limited data availability and, possibly, lack of treatment effect. However, we expect that in the (near) future, the availability of high-quality, observational data will significantly increase as advances in LLMs facilitate data re-use through data extraction and annotation. It does not seem

unreasonable to expect that this will also simplify obtaining surrogate (proxy) outcomes, and thus quicken future research. Moreover, availability of additional variables and data would enable development of more complex (Bayesian) networks and could open the door to better understanding biological processes. As such, we believe these developments will allow future research to reach more definitive conclusions.

References

1. Hendriks MP, Verbeek XAAM, van Vegchel T, et al: Transformation of the National Breast Cancer Guideline Into Data-Driven Clinical Decision Trees. JCO Clin Cancer Inform 3:1–14, 2019
2. McDonald CJ: Standards for the Electronic Transfer of Clinical Data: Progress, Promises, and the Conductor's Wand. Proc Annu Symp Comput Appl Med Care 9–14, 1990
3. De Haan W, Klein Wolterink G, van Ginneken J: Visie op zibs [Internet]. Nictiz , 2022[cited 2023 Dec 11] Available from: https://nictiz.nl/publicaties/visie-op-zibs/
4. Los R: Supporting Uniform Representation of Data. Structuring Medical Narratives for Care and Research [Internet], 2006[cited 2023 Dec 19] Available from: https://repub.eur.nl/pub/7579/060310_Los-RK.pdf
5. Gehrmann J, Herczog E, Decker S, et al: What prevents us from reusing medical real-world data in research. Sci Data 10:459, 2023
6. Ebbers T, Takes RP, Smeele LE, et al: The implementation of a multidisciplinary, electronic health record embedded care pathway to improve structured data recording and decrease electronic health record burden. Int J Med Inform 184:105344, 2024
7. Introducing ChatGPT [Internet][cited 2024 Apr 4] Available from: https://openai.com/blog/chatgpt
8. OpenAI, Achiam J, Adler S, et al: GPT-4 Technical Report [Internet], 2024[cited 2024 Apr 4] Available from: http://arxiv.org/abs/2303.08774
9. Fink MA, Bischoff A, Fink CA, et al: Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. Radiology 308:e231362, 2023
10. Sargent DJ, Wieand HS, Haller DG, et al: Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: individual patient data from 20,898 patients on 18 randomized trials. J Clin Oncol 23:8664–8670, 2005
11. Sargent D, Shi Q, Yothers G, et al: Two or three year disease-free survival (DFS) as a primary end-point in stage III adjuvant colon cancer trials with fluoropyrimidines with or without oxaliplatin or irinotecan: data from 12,676 patients from MOSAIC, X-ACT, PETACC-3, C-06, C-07 and C89803. Eur J Cancer 47:990–996, 2011
12. Ecker BL, Lee J, Saadat LV, et al: Recurrence-free survival versus overall survival as a primary endpoint for studies of resected colorectal liver metastasis: a retrospective study and meta-analysis. The Lancet Oncology 23:1332–1342, 2022
13. Greene FL, Sobin LH: The staging of cancer: a retrospective and prospective appraisal. CA Cancer J Clin 58:180–190, 2008
14. Amin MB, Edge S, Greene F, et al (eds): AJCC Cancer Staging Manual (ed 8). Springer International Publishing, 2017
15. Lu C-L, Wang S, Ji Z, et al: WebDISCO: a web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc 22:1212–1219, 2015
16. Jochems A, Deist TM, van Soest J, et al: Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital - A real life proof of concept. Radiother Oncol 121:459–467, 2016
17. Osong B, Masciocchi C, Damiani A, et al: Bayesian network structure for predicting local tumor recurrence in rectal cancer patients treated with neoadjuvant chemoradiation followed by surgery. Phys Imaging Radiat Oncol 22:1–7, 2022
18. Moncada-Torres A, Martin F, Sieswerda M, et al: VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for secure insight eXchange, in AMIA annual symposium proceedings. 2020, pp 870–877
19. Smits D, van Beusekom B, Martin F, et al: An improved infrastructure for privacy-preserving analysis of patient data, in Proceedings of the international conference of informatics, management, and technology in healthcare (ICIMTH). 2022, pp 144–147
20. Beutel DJ, Topal T, Mathur A, et al: Flower: A Friendly Federated Learning Research Framework [Internet], 2022[cited 2024 Jan 31] Available from: http://arxiv.org/abs/2007.14390

21. IDEA4RC: IDEA4RC [Internet]. Home [cited 2024 Jan 31] Available from: https://www.idea4rc.eu/idea4rc-project/
22. CODA Platform [Internet][cited 2024 Jan 31] Available from: https://www.coda-platform.com/#about
23. Mahon P, Chatzitheofilou I, Dekker A, et al: A federated learning system for precision oncology in Europe: DigiONE. Nat Med 1–4, 2024
24. Giordano SH, Kuo Y-F, Duan Z, et al: Limits of observational data in determining outcomes from cancer therapy. Cancer 112:2456–2466, 2008
25. Rubin DB, Thomas N: Matching using estimated propensity scores: relating theory to practice. Biometrics 52:249–264, 1996
26. Rubin DB: Estimating causal effects from large data sets using propensity scores. Ann Intern Med 127:757–763, 1997
27. Sakellakis M, Jacqueline Flores L, Ramachandran S: Patterns of indolence in prostate cancer (Review). Exp Ther Med 23:351, 2022
28. Bill-Axelson A, Holmberg L, Ruutu M, et al: Radical prostatectomy versus watchful waiting in early prostate cancer. N Engl J Med 352:1977–1984, 2005
29. Wong Y-N, Mitra N, Hudes G, et al: Survival associated with treatment vs observation of localized prostate cancer in elderly men. JAMA 296:2683–2693, 2006
30. Erning FN van, Janssen-Heijnen MLG, Creemers GJ, et al: Recurrence-free and overall survival among elderly stage III colon cancer patients treated with CAPOX or capecitabine monotherapy. International Journal of Cancer 140:224–233, 2017
31. Tournigand C, André T, Bonnetain F, et al: Adjuvant therapy with fluorouracil and oxaliplatin in stage II and elderly patients (between ages 70 and 75 years) with colon cancer: subgroup analyses of the Multicenter International Study of Oxaliplatin, Fluorouracil, and Leucovorin in the Adjuvant Treatment of Colon Cancer trial. J Clin Oncol 30:3353–3360, 2012
32. DeCosse JJ, Ngoi SS, Jacobson JS, et al: Gender and colorectal cancer. Eur J Cancer Prev 2:105–115, 1993
33. Marija C, Kresimir D, Ognjen B, et al: Estimation of colon cancer grade and metastatic lymph node involvement using DWI/ADC sequences. Acta Radiol 2841851221130008, 2022
34. Anglemyer A, Horvath HT, Bero L: Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials [Internet]. Cochrane Database of Systematic Reviews , 2014[cited 2023 Feb 1] Available from: https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.MR000034.pub2/full
35. Kubben P, Dumontier M, Dekker A (eds): Fundamentals of Clinical Data Science [Internet]. Cham (CH), Springer, 2019[cited 2023 Dec 19] Available from: http://www.ncbi.nlm.nih.gov/books/NBK543527/
36. Chen W, O'Bryan CM, Gorham G, et al: Barriers and enablers to implementing and using clinical decision support systems for chronic diseases: a qualitative systematic review and meta-aggregation. Implement Sci Commun 3:81, 2022
37. Meunier P-Y, Raynaud C, Guimaraes E, et al: Barriers and Facilitators to the Use of Clinical Decision Support Systems in Primary Care: A Mixed-Methods Systematic Review. Ann Fam Med 21:57–69, 2023
38. Ford E, Edelman N, Somers L, et al: Barriers and facilitators to the adoption of electronic clinical decision support systems: a qualitative interview study with UK general practitioners. BMC Med Inform Decis Mak 21:193, 2021
39. Andersen SK, Olesen KG, Jensen FV, et al: HUGIN - a Shell for Building Bayesian Belief Universes for Expert Systems. Proceedings of the Eleventh International Joint Conference on Artificial Intelligence 2:1080–1085, 1989
40. Scutari M: Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software 35:1–22, 2010
41. Ankan A, Panda A: pgmpy: Probabilistic graphical models using python, in Proceedings of the 14th python in science conference (SCIPY 2015). Citeseer, 2015

Appendices

Summary

In oncology, observational data, also known as real world data, is becoming increasingly important to monitor disease incidence and prevalence, organization and quality of care, and trends in treatment. However, both collecting and using observational data comes with challenges. Firstly, data requires a degree of standardization to make automatic collection feasible. Secondly, comparing data over longer periods of time requires that classification systems either remain unchanged, or that values can be mapped between systems and/or versions. Thirdly, data cannot always be freely shared, for example due to privacy considerations or laws and regulations. Finally, establishing causal relationships is difficult due to the potential presence of confounding bias in the data. A specific example of this type of bias that complicates estimating treatment effect is confounding by indication, also known as treatment selection bias. This thesis addresses these challenges.

In **Chapter 2** we explored the possibility of standardizing tumor board reports according to the Dutch national clinical practice guideline (CPG) for breast cancer with the intention to increase the quality of the clinical documentation and enable secondary use for registration and research. Using an information standard that was previously derived from the CPG, we developed three different tumor board forms: preoperative, postoperative, and postneoadjuvant-postoperative. Half of the items on the forms were relevant for the Netherlands Cancer Registry (NCR). The forms were implemented in the electronic healthcare record at Amphia Hospital, a large hospital in the south of the Netherlands with a focus on education, science, and innovation. We compared quality (completeness) of clinical documentation before and after implementation, as well as impact on clinical workload. Analyses showed that quality increased without impacting tumor board preparation time. As such, our work enables data reuse for secondary purposes, like cancer registries.

Subsequently, in **Chapter 3**, we show how Bayesian Networks (BNs) can be used to facilitate (unsupervised) probabilistic reclassification, enabling statistical analyses across TNM-editions in lung cancer. In oncology, the TNM classification system is used for prognosis, treatment selection, and research. Regular updates potentially break backward compatibility. Reclassification is not always possible, is labor intensive, or requires additional data. We developed a BN for reclassifying the 5th, 6th, and 7th editions of the TNM and simultaneously predicting survival for non-small-cell lung cancer (NSCLC). This was done without knowing the true TNM classification in various editions in the training set by leveraging the correlation between TNM and survival. For evaluation purposes only, part of the 7th edition test data was manually reclassified. Predicting 6th edition stage grouping using 7th edition data and vice versa resulted in average accuracies, sensitivities, and specificities between 0.85 and 0.99; the AUC for 2-year survival was 0.81.

To increase the volume of available data (or variance within data), it can be desirable to run statistical analyses across countries. However, in many cases this is hindered by privacy concerns, if not by laws and regulations. Federated Learning is a technique that solves these issues by enabling statistical analyses without sharing record level data; the data remains at the source, and only aggregated results and statistics are exchanged. In **Chapter 4** we describe how we implemented a flexible, programming language agnostic, infrastructure for Federated Learning, as well as the algorithm that calculates the Cox Proportional Hazards model. Both were employed to investigate incidence and treatment in oral cavity cancer (OCC) in the Netherlands and Taiwan. This showed that five prognostic factors (age, stage, grade, treatment modality and hospital volume) of OCC have differential effects on survival between the Netherlands and Taiwan.

In **Chapter 5** we developed a method to learn causal models from observational data. This method uses structure learning algorithms for Bayesian Networks in conjunction with (rudimentary) clinical knowledge, specifically a blacklist of known non-causal edges to guide the structure learning process. We illustrated the application of our method by an analysis of the effect of active treatment (versus observation) in localized prostate cancer. Our analysis successfully resulted in a causal model but did not identify significant association between treatment and survival. When associations between treatment and survival were manually added, graph analysis identified year of diagnosis and age as confounders; we found a treatment effect that was close to the 5-percentage point found in randomized clinical trials.

The method described above was further applied in **Chapter 6**, where we estimated the effect of adjuvant chemotherapy. While adjuvant therapy with capecitabine and oxaliplatin (CAPOX) has been proven effective in stage III colon cancer, capecitabine monotherapy (CapMono) might be equally effective in elderly patients. Unfortunately, the elderly are underrepresented in clinical trials and patients included may not be representative of the routine care population. Here, we built causal models using BN, identified confounders, and estimated the effect of adjuvant chemotherapy using survival analyses. We compared two scenarios: adjuvant treatment vs surgery only, and CAPOX vs CapMono. When comparing adjuvant treatment to surgery only, we found that adjuvant treatment was significantly associated with survival. However, when investigating CAPOX vs CapMono we did not find any association between treatment choice and survival. Analyses using Cox models did not identify significant association either. We concluded that additional oxaliplatin did not have a marked effect and should be avoided in elderly patients with stage III colon cancer.

In **Chapter 7** we once more applied the above method of developing a causal model through structure learning in conjunction with clinical knowledge. Here we investigated the possibility of creating a single BN that simultaneously predicts recurrence-free survival, overall survival, and quality of life for patients with nonmetastatic rectal cancer.

Treatment options for this disease depend on tumor stage but generally consist of surgery with or without neoadjuvant radiotherapy or chemoradiation. While neoadjuvant treatment improves local control and enables better curative options, it also incurs the risk of overtreatment and inferior functional outcome in some patients. The use of clinical decision support (CDS) systems has the potential to guide personalized care and shared decision making (SDM) by providing meaningful prognostic information. In contrast to other models (e.g., regression models), BNs do not have dedicated inputs or outputs. Instead, applying evidence calculates the conditional probabilities for all variables in the network. This, in combination with their visual representation, makes them well suited for supporting SDM. Contrary to our expectations, however, structure learning initially did not identify any associations between treatment and survival, recurrence-free survival, or PROMs. However, associations between treatment and survival *were* found if recurrence-free survival nodes were excluded from the model. By manually adding relationships to the BN, we were still able to simultaneously predict survival and PROMs with reasonable calibration.

Finally, in **Chapter 8** we revisit the major outcomes of the previous chapters and discuss limitations and implications of our work.

Samenvatting

Binnen de oncologie wordt observationele data, ook wel bekend als “real world data”, steeds belangrijker voor het toezicht houden op incidentie en prevalentie van ziekte, organisatie van zorg, kwaliteit van zorg, en trends in behandeling. Echter, zowel het verzamelen als het gebruiken van observationele data gaat met uitdagingen gepaard. Ten eerste is een zekere mate van standaardisatie noodzakelijk om het automatisch verzamelen van data mogelijk te maken. Ten tweede is het voor het vergelijken van data over langere periodes belangrijk dat gebruikte classificatiesystemen niet veranderen óf dat de codes van deze classificatiesystemen vertaald kunnen worden tussen verschillende versies. Ten derde kan data niet altijd vrijelijk gedeeld worden, bijvoorbeeld vanwege zorgen over privacy, of beperkingen door wet- en regelgeving. Ten slotte is het lastig om causale relaties te identificeren in observationele data vanwege de mogelijke aanwezigheid van confounding bias die geobserveerde relaties kan vertekenen. Een specifieke vorm van confounding die het schatten van behandeleffect lastig maakt, is confounding van de (keuze voor) behandeling of -indicatie en de uitkomst (Engels: “confounding by indication”). Dit proefschrift zoekt (en biedt) een antwoord op deze uitdagingen.

In **Hoofdstuk 2** bespreken we de mogelijkheid om de verslaglegging van het multidisciplinair overleg (MDO), een onderdeel van het oncologisch proces, te standaardiseren volgens de landelijke richtlijn voor borstkanker. Hierbij was het doel om de kwaliteit van de klinische verslaglegging te verbeteren en om hergebruik (van de data) voor registraties en onderzoek mogelijk te maken. Met behulp van een informatiestandaard, die in een eerder stadium was afgeleid van de landelijke richtlijn, ontwikkelden we drie verschillende MDO-formulieren: preoperatief, postoperatief, en postneoadjuvant-postoperatief. De helft van de items op de formulieren was relevant voor de Nederlandse Kankerregistratie (NKR). De formulieren werden geïmplementeerd in Amphia Ziekenhuis, een topklinisch ziekenhuis met een focus op innovatie in het zuiden van Nederland. We vergeleken de kwaliteit (volledigheid) van de klinische documentatie en de impact op de werkdruk voor en na implementatie. De analyse toonde aan dat de kwaliteit toenam, zonder (negatieve) invloed op de voorbereidingstijd van het MDO. Dit toont aan dat deze methode hergebruik van klinische gegevens mogelijk maakt, bijvoorbeeld voor een kankerregistratie.

Vervolgens beschrijven we in **Hoofdstuk 3** aan hoe Bayesiaanse Netwerken (BNs) gebruikt kunnen worden om data probabilistisch opnieuw te classificeren, wat het mogelijk maakt om statistische analyses uit te voeren op longkankerdata die meerdere TNM-edities/-versies overspant. Binnen de oncologie wordt het TNM-classificatiesysteem gebruikt voor het bepalen van prognose, selectie van behandeling, en (stratificeren van data in) onderzoek. Reguliere updates kunnen ervoor zorgen dat compatibiliteit met eerdere versies verloren gaat. Data opnieuw classificeren is niet altijd mogelijk, kan aanzienlijke inspanning kosten, of kan aanvullende data vergen. We ontwikkelden daarom een BN

dat het (tumor)stadium opnieuw classificeert van/naar de 5e, 6e, en 7e editie van het TNM-classificatiesysteem en tegelijkertijd overleving voorspelt voor niet-kleincellig longcarcinoom (NSCLC). Dit deden we, *zonder* dat er data beschikbaar was waarvan het tumorstadium in meerdere edities bekend was, door gebruik te maken van de correlatie tussen het tumorstadium en overleving. Voor de evaluatie van de kwaliteit van het model werd een deel van de testdata van de 7e editie handmatig opnieuw geclassificeerd in de 6e editie. Het voorspellen van het tumorstadium in de 6e editie met behulp van data uit de 7e editie en andersom leverde een gemiddelde nauwkeurigheid, sensitiviteit, en specificiteit tussen de 0.85 en 0.99; de AUC voor 2-jaarsoverleving was 0.81.

Om over voldoende data of data met meer variantie te beschikken, kan het nodig zijn om onderzoek internationaal uit te voeren. In veel gevallen stuit dit echter op bezwaren vanwege (zorgen over) privacy, en wet- en regelgeving. Gefedereerd leren (Engels: Federated Learning) is een techniek die deze problemen oplost door statistische analyses met meerdere partijen mogelijk te maken zonder gegevens op patiëntniveau te delen; de data blijft bij de bron en alleen geaggregeerde resultaten en statistieken worden uitgewisseld. In **Hoofdstuk 4** beschrijven we hoe we een flexibele infrastructuur voor gefedereerd leren hebben ontwikkeld die programmeertaal agnostisch is. Daarnaast beschrijven we de implementatie van het Cox Proportional Hazards model voor deze infrastructuur. Beide werden ingezet voor onderzoek naar incidentie en behandeling van mondholtecarcinoom in Nederland en Taiwan. Hieruit bleek dat vijf voorspellende factoren (leeftijd, stadium, graad, behandelmodaliteit, en ziekenhuisvolume) een verschillend effect hadden op overleving tussen Nederland en Taiwan.

In **Hoofdstuk 5** hebben we een methode ontwikkeld om causale modellen te "leren" uit observationele data. Hierbij werden bestaande algoritmes voor BNs, die een netwerkstructuur tussen variabelen kunnen afleiden uit data (Engels: structure learning), gecombineerd met (beperkte) klinische kennis. Deze kennis heeft de vorm van een lijst van netwerkverbindingen waarvan bekend is dat een bepaalde richting van causaliteit onmogelijk is. Bijvoorbeeld, een uitkomst zal nooit de *oorzaak* zijn van een behandeling; deze kennis kan worden toegevoegd aan de lijst. Aan de hand van een analyse van het effect van actieve behandeling (versus afwachten) bij gelokaliseerd prostaatcarcinoom lieten we zien hoe onze methode toegepast kan worden. Hierbij werd met succes een causaal model ontwikkeld maar geen (statistisch) significante associatie gevonden tussen behandeling en overleving. Nadat verbindingen tussen behandeling en overleving handmatig waren toegevoegd, identificeerde een analyse van het netwerk de variabelen jaar van diagnose en leeftijd als confounders, en vonden we een behandeleffect dat dicht in de buurt lag bij de 5-procentpunt die gevonden wordt in gerandomiseerde klinische trials.

Bovenstaande methode werd nogmaals toegepast in **Hoofstuk 6**, waar we het effect hebben geschat van adjuvante chemotherapie bij coloncarcinoom. Hoewel adjuvante

therapie met capecitabine en oxaliplatin (CAPOX) bewezen effectief is in stadium III coloncarcinoom, bestaat de mogelijkheid dat capecitabine monotherapie (CapMono) in oudere patiënten even effectief is. Helaas zijn oudere patiënten ondervertegenwoordigd in klinische trials en zijn geïncludeerde patiënten niet altijd representatief voor de populatie in de dagelijkse praktijk. Daarom hebben we met behulp van BNs causale modellen ontwikkeld, confounders geïdentificeerd, en het effect van adjuvante behandeling geschat met overlevingsanalyses. We vergeleken hierbij twee scenario's: adjuvante behandeling ten opzichte van alleen chirurgie, en CAPOX ten opzichte van CapMono. Bij de vergelijking tussen adjuvante behandeling met alleen chirurgie vonden we een significante associatie met overleving. Echter, bij de vergelijking tussen CAPOX met CapMono vonden we geen associatie tussen keuze van behandeling en overleving. Ook analyses met behulp van Cox-modellen vonden geen significante associatie. We concludeerden daarom dat aanvullende toediening van oxaliplatin geen effect heeft op overleving en vermeden zou moeten worden bij oudere patiënten met stadium III coloncarcinoom.

In **Hoofdstuk 7** gebruikten we bovenstaande methode voor het leren van een causaal model om te onderzoeken of het mogelijk is om een enkel BN te creëren dat tegelijkertijd algemene overleving, recidiefvrije overleving, en kwaliteit van leven voorspelt voor patiënten met niet-gemetastaseerd rectumcarcinoom. De behandeling van deze ziekte is afhankelijk van het tumorstadium, maar bestaat over het algemeen uit chirurgie, met of zonder neoadjuvante radiotherapie of neoadjuvante chemoradiatie. Hoewel neoadjuvante behandeling de lokale controle verbetert en kan leiden tot betere curatieve opties, introduceert dit ook het risico op overbehandeling en slechtere functionele uitkomsten. De inzet van elektronische beslissingsondersteuning in de kliniek (Engels: Clinical Decision Support; CDS) heeft, doordat dit betekenisvolle prognostische informatie kan verschaffen, de potentie om richting te geven aan gepersonaliseerde zorg en samen beslissen (Engels: shared decision making; SDM). In tegenstelling tot andere soorten modellen (bijvoorbeeld regressiemodellen) hebben BNs geen vaste variabelen voor in- of uitvoer. In plaats daarvan kan het netwerk worden gevoed met bewijs (over de status van één of meer variabelen) waarna de conditionele kansen voor alle overige variabelen worden uitgerekend. Dit, in combinatie met hun visuele weergave, maakt BNs erg geschikt voor het ondersteunen van SDM. In tegenstelling tot onze verwachtingen vonden we echter geen associaties tussen behandeling en algemene overleving, recidiefvrije overleving, of kwaliteit van leven. Als de variabelen met betrekking op recidiefvrije overleving uit het model werden gelaten, werden associaties tussen behandeling en algemene overleving wél gevonden. Door handmatig relaties aan het model toe te voegen, waren we toch in staat om tegelijkertijd overleving en kwaliteit van leven te voorspellen, waarbij de kalibratie van de voorspelde kansen acceptabel was.

Research Impact

Introduction

In oncology, observational data, also known as real world data, is becoming increasingly important to monitor disease incidence and prevalence, organization and quality of care, and trends in treatment. However, both collecting and using observational data comes with challenges. Firstly, data requires a degree of standardization to make automatic collection feasible. Secondly, comparing data over longer periods of time requires that classification systems either remain unchanged, or that values can be mapped between systems and/or versions. Thirdly, data cannot always be freely shared, for example due to privacy considerations or laws and regulations. Finally, establishing causal relationships is difficult due to the potential presence of confounding bias in the data. A specific example of this type of bias is confounding by indication, also known as treatment selection bias.

This research explored new possibilities for increasing the potential of observational data by investigating four questions: how can we 1) better extract data from the clinical process without negatively impacting it, 2) ensure that historically collected data retains its value, 3) reuse data across (international) organizations safely and without infringing on patient privacy, and 4) broaden the set of questions we can answer to include estimation and comparison of treatment effect.

Extracting data from the clinical process

This thesis shows that, through standardization, it is possible to both improve the quality of clinical documentation for tumor board meetings, and to simultaneously enable secondary use of said documentation. Tumor board meetings, pivotal moments in the clinical process, were chosen because they act as convergence points that are clearly marked within the process and align with registry interests. As such, generalizability to other clinical processes hinges on the level of their standardization and availability of Clinical Practice Guidelines. And while extension to other tumor types is feasible, not all medical processes have such clearly identifiable convergence points.

In many cases, improving healthcare begins with gaining insight from current clinical practice, which requires structured data for analysis. Standardization of clinical documentation allows for case-analysis at a single site, or between sites. This research contradicts a long-standing myth that structured data capture, increases physician workload. As such this knowledge should be useful for healthcare professionals looking to improve their documentation and facilitate research, as well as for cancer registries that depend on clinical data. Ideally, both to improve cancer care and to further research, countries should develop standardized tumor board report forms. If possible, these would be incorporated into the clinical practice guidelines. In the Netherlands, the Netherlands Comprehensive Cancer Organisation (IKNL), supports this work.

Ensuring data retains its value

Furthermore, this thesis has shown how unsupervised probabilistic reclassification can be used to retain compatibility between two editions of the TNM-staging system. This was done without using a training set with labels for both versions, but by leveraging the association between TNM-stage and survival instead. Applying this method requires limited (clinical) knowledge about relationships between source and target system and either a common cause or effect. However, the biggest obstacle for more general use, is that the evaluation of the resulting probabilistic mapping requires a test set, and thus data that is classified in both source and target classification system. If it is possible to obtain a test set, it might also be possible to use (part of) this dataset for training purposes.

The most obvious benefit of the ability to reclassify old data into newer classification systems (or vice versa), is that this enables analyses spanning longer time periods. This is especially helpful for organizations that deal with longer time spans, such as cancer registries. Here, this technique could help avoid "breaks" in the data otherwise caused by changes to the underlying TNM-classification. However, practical adoption is likely to require development of easy-to-use software or to incorporate this method into existing statistical packages.

Using data across borders

If data from a single site or country is insufficient, Federated Learning facilitates statistical analyses without sharing record level data. This unlocks data for reuse between organizations without infringing on patient privacy and without relinquishing control of the data, even across borders. The technique is broadly applicable, although it requires data to be standardized across participants. The software that was started as part of this thesis was published as open source and has since evolved into a mature application known as vantage6.

Federated Learning enables analyses spanning more sites/larger geographical areas since it alleviates privacy concerns. Additionally, it addresses frequently observed concerns regarding data sharing. For example, it ensures data providers remain in control of the data and mitigates the risk of data leaks. This makes the technique well suited for international research, as illustrated by projects like IDEA4RC which focuses aims to develop an IT infrastructure to facilitate the sharing and re-use of health data among (European) clinical centers to promote research on rare cancers. International organizations, such as the European Cancer Organization, might use Federated Learning to (virtually) combine information across nations.

Broadening the use of observational data

Finally, this thesis describes a method, leveraging structure learning algorithms for Bayesian Networks, to create causal models that can be used to identify confounders and estimate treatment effect. This allows analyses based on routinely collected clinical data that previously required data from Randomized Controlled Trials (RCTs). Still, the method has a few limitations. For instance, it depends on the ability to specify non-causal relationships in the data, which is not always possible. Additionally, the faith in the resulting model depends on (clinical knowledge to) trust that there is no unmeasured confounding.

The ability to reliably identify confounders and estimate treatment effect provides an important step forward in the application of observational data. Most importantly, it may drastically reduce the turnaround time between hypothesis and answer. Generally, clinical research that attempts to estimate treatment effect (or compare treatments) requires an RCT. If the endpoint of the trial is 5-year survival, results cannot be expected any sooner than five years after the start. When observational data can be used to answer the same question, this 5-year delay may be circumvented. As such, this thesis provides a method to improve patient care by learning from readily available, real-world data and without the delays that are typically associated with RCTs. Although improving patient care may be the ultimate goal, the method developed in this thesis is most likely of interest to (clinical) researchers. As such, practical adoption would benefit from development of easy-to-use software or to incorporate this method into existing statistical packages.

Curriculum Vitae

Melle Sieswerda was born on June 9th 1980 in Apeldoorn, the Netherlands. He grew up in Arnhem and later moved to Didam. He finished his secondary education at Stedelijk Gymnasium Arnhem and went on to study Computer Science at Delft University of Technology in 1999.

In 2004, just before enrolling in the Bioinformatics master programme in Delft, he thought that having skills in both computer science and medicine would be an interesting and potential valuable combination. Therefore, he enrolled at the Leiden University (Medical Center) to study Medicine in parallel. Here, he met Kawita, his future wife.

After finishing his studies in Bioinformatics by writing & defending his master's thesis on the subject of integrating different microarray datatypes in 2007, he decided to start a professional career at Siemens Healthcare. This was also to support his continuing education in Medicine. The intention was to split work and study 80/20; in practice it proved difficult to find time to study.

Still, in 2012 the clinical rotations (Dutch: coschappen), the internships that make up the final part of the curriculum for Medicine, came into view. Since these are generally a fulltime endeavor, he left Siemens after 4,5 years. Completing clinical rotations took a little longer than usual, due to the birth of his first daughter, Varisha, early 2013. However, these were successfully completed in 2014.

Next, in 2015 he started at the Netherlands Comprehensive Cancer Organisation (better known as IKNL) as a Clinical Informatician. Here he worked at bridging the gap between healthcare and IT through projects that focused on standardizing clinical documentation. To maintain his clinical skills he, additionally started as a physician at the Leiden University Medical Center (LUMC) in the field of Palliative Care. In 2016 his son, Anil, was born.

To focus more on research, he transitioned when a new department of Clinical Data Science was created. Here, he worked on creating additional value for the Netherlands Cancer Registry through federated learning and causal inference. This led to the opportunity of starting a PhD at Maastro Clinic, resulting in this thesis.

List of Publications

Menezes RX, Boetzer M, Sieswerda M, van Ommen GJB, Boer JM. Integrated analysis of DNA copy number and gene expression microarray data using gene sets. *BMC Bioinformatics*. 2009;10:203. doi:10.1186/1471-2105-10-203

Ebben KCWJ, Sieswerda MS, Luiten EJT, et al. Impact on Quality of Documentation and Workload of the Introduction of a National Information Standard for Tumor Board Reporting. *JCO Clin Cancer Inform*. 2020;4:346-356. doi:10.1200/CCI.19.00050

Geleijnse G, Chiang RCJ, Sieswerda M, et al. Prognostic factors analysis for oral cavity cancer survival in the Netherlands and Taiwan using a privacy-preserving federated infrastructure. *Sci Rep*. 2020;10(1):20526. doi:10.1038/s41598-020-77476-2

Moncada-Torres A, Martin F, Sieswerda M, van Soest J, Geleijnse G. VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for secure insight eXchange. In: *AMIA Annual Symposium Proceedings*. ; 2020:870-877.

Sieswerda MS, Bermejo I, Geleijnse G, et al. Predicting Lung Cancer Survival Using Probabilistic Reclassification of TNM Editions With a Bayesian Network. *JCO Clin Cancer Inform*. 2020;4:436-443. doi:10.1200/CCI.19.00136

van der Stap L, van Haaften MF, van Marrewijk EF, et al. The feasibility of a Bayesian network model to assess the probability of simultaneous symptoms in patients with advanced cancer. *Sci Rep*. 2022;12(1):22295. doi:10.1038/s41598-022-26342-4

Sieswerda M, van Rossum R, Bermejo I, et al. Estimating Treatment Effect of Adjuvant Chemotherapy in Elderly Patients With Stage III Colon Cancer Using Bayesian Networks. *JCO Clinical Cancer Informatics*. 2023;(7):e2300080. doi:10.1200/CCI.23.00080

Sieswerda M, Xie S, van Rossum R, et al. Identifying Confounders Using Bayesian Networks and Estimating Treatment Effect in Prostate Cancer With Observational Data. *JCO Clin Cancer Inform*. 2023;7:e2200080. doi:10.1200/CCI.22.00080

Acknowledgments

First of all, I would like to thank the members of the assessment committee. The time and effort they put into reading this dissertation is greatly appreciated.

Dear André and Inigo, without your help and encouragement, this endeavor would likely not have borne fruit. The distance between Oegstgeest and Maastricht was a stretch, but I truly enjoyed our discussions about the methodology underlying so much of this thesis. Finding explanations together for the results (network structures) we obtained was challenging and at the same time rewarding when a follow-up analysis corroborated our hypothesis. Furthermore, I owe you a great debt of gratitude for all contributions to this dissertation and the support you have provided over the past years.

Dear Xander, without your creative and ambitious spirit, and equally strong drive for improving healthcare, I probably would not have joined IKNL. Also, without your focus on data driven healthcare, the department of Clinical Data Science would not have been created, and you would not have suggested I started a PhD. You were the spark that started this work. I am tremendously grateful for all the opportunities you have provided, and the faith you placed in me.

Dear Valery, although you were less involved with the day-to-day activities, your support meant a lot. Your presence in the background provided reassurance when needed.

Dear Gijs, as head of the Clinical Data Science group at IKNL you provided a pleasant working environment, giving me room to experiment and grow. You were the big brother I could look up to at work. Additionally, you actively contributed to many of the manuscripts that found their way into this dissertation. Here too, I truly appreciate our discussions and collaboration.

Furthermore, I would like to thank my colleagues at IKNL for their support and being fantastic people to work with. In no particular order: Jurrian, Peter, Martin, Bas, Kees, Thijs, Willem, Frank, Bart, Arturo, Anja, Harm, Eva, Maarten (x2), Dimitris, Maaike, and Harmke. And of course, the registration team of the Netherlands Cancer Registry. They are too many to name individually, but they do invaluable work in collecting data; each day they contribute to IKNL's mission of reducing the impact of cancer and without their efforts none of the publications in this dissertation would have been possible.

Additionally, I would like to thank my parents, Kees and Winnie, for their support and dedication. Your continued interest helped tremendously in bringing this dissertation to a conclusion. At some point I even got the impression that you cared more about this effort than I did myself.

Of course, my brother Jouke and friend-since-secondary school Piet will be indispensable as paranymphs during the public defense of this thesis. I really appreciate that you will be by my side during this memorable occasion.

And finally, I would like to thank Kawita, Varisha, and Anil, my wife and children. Kawita, many times you patiently listened to my rambling about analyses that did not return the results I was expecting. Talking to you helped me develop new viewpoints. Also, your practical support in ensuring I had time to write, as well as your encouragement were indispensable. Varisha and Anil, your patience, support, and hugs were equally valuable, and put every struggle into perspective.

For those of you who feel left out: my sincerest apologies. It certainly was not my intention to forget anyone but given how my brain works, I'm not surprised if I did. Let me know and I will make amends in the second edition.