

Report and dissemination of findings on ZonMw project 10430252210006

The Effect of the Covid-19 Pandemic on Cancer Screening and Cancer-Related Mortality in The Netherlands in 2020 and 2021

November 2023

Marc Jacobs¹ and Ronald Meester²

¹ “Modelling and Statistics for You”; MSJAdvies@gmail.com.

² Department of Mathematics, Vrije Universiteit Amsterdam; r.w.j.meester@vu.nl.

Contents

Summary.....	5
Background	5
Introduction.....	6
Original objective & research questions.....	7
Population Screening	8
Cervical cancer screening.....	8
Colorectal cancer screening.....	8
Breast cancer screening	9
Data sources envisioned in our original proposal	9
Availability and use of open data in answering our research questions	10
Road to CBS data.....	11
Road to IKNL data.....	11
Acceptance from ZonMw.....	12
Synthetic dataset.....	12
NKR dataset	13
IKNL judgement form.....	13
Back to the original plan with more focus and four scenarios to explore	14
Methods.....	15
Components of our study and definitions used.....	15
Datasets used	17
CBS.....	17
Eurostat	18
IKNL	18
OECD	18
OpenDIS.....	19
Our World in Data.....	20
State of Public Health and Health Care.....	21
VEKTIS.....	21
VZinfo	21
Human Mortality Database	22
Building a clinical prediction model.....	22
Rapid review of clinical prediction models for breast cancer.....	23
Analysis plan of our previous work	24
Analysis plan of our current work	26
What can we say at the end of all of this?.....	26

Rapid review on the possible influence of delayed care during Covid-19 on breast, cervical and colon cancer mortality.....	27
Clinical expertise.....	28
Results.....	28
Life expectancy.....	29
Mortality.....	30
Excess mortality.....	34
Cause of death.....	42
Covid-19	48
Cancer	51
Cancer incidence and prevalence	62
Cancer screening	69
Cancer survival	72
Healthcare expenditure	76
Clinical prediction model for breast cancer.....	98
Rapid review of clinical prediction models for breast cancer.....	98
Results of our previous work.....	99
Results of our current work.....	100
Validity of the IKNL Synthetic data	112
Most common groups in breast cancer IKNL synthetic data	116
Clinical prediction models found.....	128
Comparison of clinical prediction founds and the clinical prediction model built.....	129
Literature research on the influence of delayed care during Covid-19 on breast, cervical and colon cancer mortality.....	130
Clinical expert opinion.....	130
Jan Bonte (neurologist).....	131
Roderik Kraaijenhagen (cardiologist) and Sabine Pinedo (internist)	133
Discussion	134
Data landscape in the Netherlands.....	135
Four possible scenarios.....	136
There is no excess mortality in 2020 and 2021	137
The excess mortality from 2020 and 2021 can be fully explained by Covid-19.....	141
The excess mortality from 2020 and 2021 can be (partly) explained by the temporary cessation of population screening.....	147
The excess mortality from 2020 and 2021 can be (partly) explained by delayed care of already diagnosed cancer patients	150
Limitations of our research.....	150

Directions for future research	152
Final thoughts.....	152

Summary

This report would have been different had we been given access to the patient-level data we deemed useful to provide more robust inferences. Nevertheless, in the absence of such data and the presence of only openly available data, it seems highly unlikely that the excess mortality seen in 2020 and 2021 is due to delayed health care. Not only are there no hints for such a mechanism in the literature, but there is also no indication for it at all in the data openly available, as we have shown in detail. We did not observe any serious trend reversal apart from the decrease in screening rates and cancer diagnoses. This means that the cause for the excess mortality seen in 2021 and beyond (which is higher than the Covid-19 mortality) must be found elsewhere.

Background

We find it important to describe the background of how this study came about. This study was sponsored by ZonMw³ which is an organization tasked with programming and distributing funds from the Dutch Government for health research and innovation in care.⁴ ZonMw has as its main clients the Ministry of Health, Welfare and Sport, and NWO.

In June 2022, ZonMw released a document showing their research agenda⁵ on investigating the excess mortality in the Netherlands for the period 2020 - 2021. This agenda was a direct result of motion 1617 (initiated by Pieter Omtzigt) which was unanimously (!) accepted by the Dutch parliament on December 2, 2021.⁶ The motion calls for the earliest possible academic research into the causes of excess mortality during the period March 2020 to November 2021. A guidance committee was put in place to develop a research agenda which was categorized into three lines:

1. Systematic literature analysis with international comparison.
2. Research with available data from existing (research) cohorts; routinely collected healthcare data and registrations such as those from the Central Bureau of Statistics (CBS).

³ <https://www.zonmw.nl/en>

⁴ <https://www.zonmw.nl/en/about-zonmw>

⁵ https://www.zonmw.nl/sites/zonmw/files/2023-01/onderzoeksagenda-oversterfte_rapport_def.pdf

⁶ <https://www.tweedekeamer.nl/kamerstukken/moties/detail?id=2021Z22246&did=2021D47333>

3. Research with data from RIVM, GGD GHOR and GGDs, such as vaccination and testing data.

Our research proposal⁷ fits within the second line and was awarded funding on December 13, 2022 alongside ten other projects.⁸ In short, we focused on the effect of the Covid-19 pandemic on cancer screening and cancer-related mortality in The Netherlands in 2020 and 2021.

Introduction

This report shows our rationale, methods, and findings which we will discuss at the end. Because study results always depend on the questions asked, methods used and people involved, we will try to provide a complete picture of our endeavor. To support this we have created a GitHub⁹ page which is publicly available and can be found [here](#). On this page, we have added noteworthy documentation, the data sets used, [codes applied](#), and results obtained.

It is important to mention at an early stage that our road was much more difficult than we had envisioned. Although nobody can realistically anticipate all possible roadblocks when drafting a research proposal, it surprised us how difficult it was to obtain data for a granted research proposal trying to answer an important societal issue. Hence, the content of our report has (unintentionally) become twofold:

1. To describe the road and setbacks we faced when requesting the data we considered useful for meeting our objective. We will therefore highlight our experience of the current data landscape of the Netherlands as far as epidemiological data is concerned.
2. To show, in a transparent way, our methods used and results obtained.

To provide a report that is transparent and clear, we have chosen to use the [introduction](#) section to first describe the [original objective and research questions](#), as shown in the research proposal, accompanied by some background epidemiology. This should place our study in its original context. We will then describe [the road towards obtaining the necessary CBS data](#) for answering these questions which turned out to be extremely difficult and costly.

⁷ <https://projecten.zonmw.nl/nl/project/effect-delayed-cancer-screening-stage-shift-and-increased-mortality-during-coronapandemic>

⁸ <https://www.zonmw.nl/nl/onderzoek-naar-redenen-en-oorzaken-van-oversterfte-de-coronacrisis>

⁹ <https://github.com/>

In the second part of the introduction section, we will describe [our request to ZonMw for an extension](#), change of data source, and additional funding. It is here that we will shift from ecological data [to obtaining patient-level data from the Dutch Cancer Registry](#). In the end, this turned out to be an unsuccessful road.

In the third and final section, we will describe the final road taken towards meeting our objective and answering our research questions. Because of the limited data availability we chose to focus on the second research question, which was in essence the main objective of our research. This road leads to the [Methods section](#) which of course determines the [Results section](#) and fuels the [Discussion section](#). It is here where we will reflect on our choices made, our eventual findings, and the limitations of our research.

Original objective & research questions

The objective of this project, as described in the application form, was the following:

“[...] to estimate the effect of delayed cancer screening during the pandemic mortality for the three cancers for which screening exists in The Netherlands: breast, cervical and colorectal cancer on (excess) mortality. We will use 2019-2021 claims data to predict the probability of screening within the eligible population and mortality. Data will be analysed using a two-stage residual inclusion model where Covid-19 in 2020 will be used to control for unobserved factors influencing both screening and mortality. Our study will analyse the proportion of deaths in 2021 caused by changes in cancer mortality due to the change in screening rates.”

Our main hypothesis was that patients were more likely to die of cancer in 2021 than in the years prior due to delayed and suspended cancer screening due to the Covid-19 pandemic. From this objective and hypothesis, we deduced three research objectives we considered necessary to answer in providing an answer to the objective as stated above:

1. *Can we develop algorithms to identify metastatic cancer in the OpenDIS claims data?*
2. *What is the effect of the Covid-19 pandemic on cancer screening in The Netherlands and how did this impact (excess) mortality in 2021?*
3. *What underlying factors explain reasons for non-attendance in cervical, colorectal and breast cancer screening?*

Before going into some more depth into our original objective and research questions, we believe it to be of value to provide some background information on population screening in general, and some specifics about the epidemiology of cervical, colon and breast cancer.

Population Screening

Each year, many [indicators¹⁰](#) are collected to monitor and evaluate population screening¹¹. Annually, the monitoring party provides a report containing the main outcomes of the population study compared to previous years. This report, called the Monitor, includes data on participation, the number of referrals and the number of abnormalities found. The monitors of the past 5 years can be found [here](#).¹² Each year the figures are updated so that the most recent monitor shows the current results of the past years. The purpose of the monitor is to monitor the quality of population screening for cervical cancer and to identify important trends. The results are based on a defined set of indicators. The monitoring is carried out using data from ScreenIT, the national information system for population-based cancer screening, and the Pathological Anatomical National Automated Archive (PALGA). In addition, information on cancer mortality comes from CBS and information on cancer incidence and interval cancers from the Netherlands Cancer Registry (NKR).

Cervical cancer screening

To detect (pre-stages of) cervical cancer at the earliest possible stage, the government has offered nationwide population screening for cervical cancer (secondary prevention) since 1996.¹³ For cervical cancer screening, the eligible population in The Netherlands consists of women between ages of 30 and 60; approximately 4 million women.¹⁴ They receive an invitation for smear tests every five years. A full screening round lasts five years and participation is surveyed at three points in time (15, 27 and 60 months after the start of the invitation year). The five-year coverage declined from 2017 to 2021. 2020 saw the largest decrease. In that year, due to the temporary suspension of the population screening due to the COVID-19 outbreak, fewer women were invited and participated. In 2021, a total of 555,515 women participated in the population screening for cervical cancer. This corresponds to 54.8% of women invited in that year. Of the invited women, 42.7% participated by smear and 12.1% by self-sampling. In 9.5% of the participants, the high-risk Human Papilloma Virus (HPV) was found and 2.7% of the participants were directly referred to the gynaecologist for diagnostic examination. In 6,246 women there was a diagnosis of a preliminary stage of cervical cancer. This corresponds to 1.1% of population screening participants.

Colorectal cancer screening

For colorectal cancer, the eligible population is men and women between ages 55 and 75. This is approximately 4 million people of which 71.6% participated in screening in 2020. In

¹⁰ [Indicatoren | RIVM](#)

¹¹ [Monitoring en evaluatie | RIVM](#)

¹² <https://www.rivm.nl/bevolkingsonderzoek-baar-moederhalskanker/professionals/monitoring-en-evaluatie>

¹³ [Baarmoederhalskanker | Preventie | Volksgezondheid en Zorg \(vzinfo.nl\)](#)

¹⁴ [Mannen en vrouwen \(cbs.nl\)](#)

2021, 70.6% of the eligible population underwent screening.¹⁵ The number of men and women who received invitations for population screening for bowel cancer in 2021 was 2,312,606.¹⁶ The number of men and women who participated in the population-based bowel cancer screening in 2021 was 1,632,493. The number of clients referred for follow-up hospital screening in 2021 after participating in the population-based bowel cancer screening was 74,309. This is 4.6%. A total of 2,790 colon cancers and 16,878 colon cancer precursors were found in the participants invited in 2021.

Breast cancer screening

Approximately 3 million women are eligible for mammography between the ages 50 and 75. In 2020, 70.4% of women received mammography, where this number was 72.5% in 2021.¹⁷ In 2021, over 1,221,792 women in the Netherlands received invitations for the population screening for breast cancer. Besides the measures due to COVID-19, staff shortages at screening also played a role in the delay in invitations. In 2021, 885,937 women took part in the population screening for breast cancer. This is 72.5% of all women invited. The attendance rate may be lower in 2021 due to fear of COVID-19, despite the precautions taken in the screening units that were necessary to keep a distance. There are 71 screening centers in the Netherlands: 58 mobile and 13 fixed. Through the mobile screening centers, we aim to offer screening as close to home as possible. Of every 1,000 women who take part in the screening, 26 are referred for further examination in hospital. The population screening detected 6,362 cases of breast cancer in 2021.¹⁸

Data sources envisioned in our original proposal

In our original proposal, we wrote down the following with regards to the data sources envisioned:

"For our study, we will use claims data from OpenDIS which contains information about Diagnosis-Treatment Combinations (DBC) originating from the DBC information system (DIS) of the Dutch Healthcare Authority (Nza)¹⁹. We will use data for the years 2019, 2020 and 2021. The Central Bureau of Statistics (CBS) also has DBC-MSZ data from Vektis. This concerns the same data, but the Vektis data are more complete in terms of numbers of DBCs than the Nza data. DBC-MSZ also contains outpatient clinics and all MSZ institutions and it contains data on all performed procedures. We will request access to both OpenDIS and DBCMSZ where the 2020 data should be available for remote access and the DBCMSZ 2021 data is expected in Q1 of 2023."

¹⁵ [Darmkanker: deelname bevolkingsonderzoek | De Staat van Volksgezondheid en Zorg \(staatvenz.nl\)](https://www.bevolkingsonderzoeknederland.nl/borrelen-effectief/#home)

¹⁶ <https://www.bevolkingsonderzoeknederland.nl/borrelen-effectief/#home>

¹⁷ [Borstkanker: deelname bevolkingsonderzoek | De Staat van Volksgezondheid en Zorg \(staatvenz.nl\)](https://www.bevolkingsonderzoeknederland.nl/borrelen-effectief/#home)

¹⁸ <https://www.bevolkingsonderzoeknederland.nl/borrelen-effectief/#home>

¹⁹ Nederlandse zorgautoriteit. OpenDIS 2022. Accessible at: <https://opendisdata.nl/>

These data sources would provide the necessary information to tackle the four main components of our study:

1. Information on incidence and prevalence of excess mortality.
2. Information on incidence and prevalence of population screening.
3. Information on incidence and prevalence of disease.
4. Information on incidence and prevalence of disease care.

All sources are ecological sources which would mean that we would only be able to connect these four major variables on an ecological level and look for trend reversals.

Availability and use of open data in answering our research questions

From the start we envisioned the use of open data. Not only does the Dutch Government promote the availability of open data, but availability is also warranted by Dutch law to the level that the data is anonymous. This means that summary data should be available to every Dutch citizen, and we will later show that much is indeed open. In the text below, we describe the envisioned relationship between each of the datasets and our original research questions:

- a. *Can we develop algorithms to identify metastatic cancer in the OpenDIS claims data?*
[OpenDIS](#) contains data on all treatment trajectories in specialty care, from 2012 to the present. From OpenDIS we would develop algorithms for each cancer type using variables from the claims, including diagnoses, procedures, drugs, and oncologist visits.
- b. *What is the effect of the Covid-19 pandemic on cancer screening in The Netherlands and how did this impact (excess) mortality in 2021?*
To answer this question, we would estimate and compare models using estimates of 2019 screenings to predict the probability of death in 2020. We would then compare this estimate to models estimating the 2020 cancer screenings to predict the probability of death in 2021. By comparing the screened population to the eligible population that did not get screened, we would be able to control for factors other than the Covid-19 pandemic.
- c. *What underlying factors explain reasons for non-attendance in cervical, colorectal and breast cancer screening.*
Previous studies focused exclusively on the screened population, but there is a body of literature on reasons for non-attendance and non-adherence to screening programs that may help explain other factors affecting screening rates. For this, we would conduct a literature search to help guide the quantitative analysis.

Road to CBS data

We already mentioned that to answer our research questions, we would make a request to [CBS](#) to provide us with the necessary access to their secured environment. By connecting open data with CBS data, we would obtain not only more accurate data but also data beyond the ecological level. Unfortunately, this turned out to be a much more difficult road than envisioned, which we also mentioned in public.^{20 21} Despite sounding plausible at first²² it was deemed too difficult²³ or even impossible to retrieve the necessary data.²⁴ Hence, we had to find a new road towards obtaining patient-level information. In addition, because of the large amount of time lost in a project with limited time in the first place, we decided to focus on our second research question.

We will now describe our road to [IKNL](#) data. We believe this to be of interest to the Dutch public because of the false general notion that important data is easily available to independent researchers. By describing our process in full, we want to show that this is not the case. In addition, we will show that by not being able to obtain the IKNL data, we lost our single-best method of answering the objective of our study.

Road to IKNL data

The Netherlands comprehensive cancer organization (IKNL) is the most prominent quality institute for oncological and palliative research and practice.²⁵ In essence it is responsible for combining all the cancer data collected in the various hospitals that offer cancer care. This allows further research and quality control and has led to the Netherlands Cancer Registry (NCR).²⁶ Within this registry there is data since 1989 offering statistics on cancer and cancer care in the Netherlands. The IKNL is the maintainer of that registry and can decide who gets access to the data. Some data is free for all to see such as numbers on incidence, prevalence, survival, and mortality.²⁷ However, for the purpose of this study, we wanted to have access to patient-level data for which we filed a data request.²⁸ IKNL only provides data from the NCR if it is used for scientific research or statistics. Also, the purpose of the application must match the objectives of IKNL: improving oncological and palliative care or supporting scientific research and statistics.

²⁰ <https://www.volkskrant.nl/columns-opinie/opinie-regel-direct-toegang-tot-medische-gegevens-voornoafhankelijk-onderzoek-naar-oversterfte~b5083299/>

²¹ <https://nos.nl/nieuwsuur/artikel/2439513-vervolgonderzoek-naar-oversterfte-moeilijk-door-beperkte-toegang-data>

²² <https://www.trouw.nl/binnenland/onderzoekers-krijgen-waarschijnlijk-toegang-tot-cbs-vaccinatiecijfers~b36270a4/>

²³ <https://www.tweedeekamer.nl/nieuws/kamernieuws/gesprek-over-data-voor-onderzoek-naar-oversterfte>

²⁴ <https://www.mariannezwagerman.nl/eline-van-den-broek-juridische-procedure-om-oversterfte-data-in-handen-te-krijgen-nieuws/>

²⁵ <https://iknl.nl/en/about-iknl>

²⁶ <https://iknl.nl/en/ncr>

²⁷ <https://iknl.nl/nkr-cijfers?lang%7Clanguage=en>

²⁸ <https://iknl.nl/en/ncr/apply-for-data>

Acceptance from ZonMw

First, we recognized that trying to obtain IKNL data was outside our original method, and so we made a request to ZonMw to ask for more time, and additional funding, to allow us to obtain the IKNL data. The original request can be found [here](#), and the response [here](#). In summary, the reply from ZonMw was positive in supporting our endeavor to obtain patient-level NKR data from IKNL. In addition, we received a new end-date for a project which was an extension of about three months (deadline October 13, 2023). We immediately initiated communications with IKNL by sending a submission form.

Synthetic dataset

Before we describe our road to the IKNL-NKR data in full, it is noteworthy to mention that IKNL has invested considerable time in producing synthetic data.²⁹ Synthetic data is data that is not collected, and thus cannot be traced back to the original person. What it does contain, however, is all the connections found in the original dataset. Hence, when analyzing a synthetic dataset for the relationship between age and mortality, it should offer the same relationship (in direction and magnitude, but not necessarily the exact number) as when analyzing the patient dataset. In summary, the synthetic dataset could or should be useful to build a clinical prediction model (which is part of research question two) as it contains the original correlation and covariance matrixes. However, noteworthy limitations are:

1. It only contains data up until 2019.
2. It does not have all the variables included that the NKR data possesses.
3. It is only available for breast cancer.
4. It most likely does not accurately represent the patient dataset.

Point 4 needs some additional argumentation. In a [separate endeavor](#), which was unrelated to this submission, one of the authors (MJ) already looked at the value of the synthetic dataset.³⁰ The synthetic dataset showed some strange relationships between clinical variables for which the precise relationship is known. It seemed as if the covariance matrices did not fully translate, and these findings were discussed with IKNL in an online meeting upon showing the results.

Since IKNL is (rightfully) cautious in their use of their synthetic data, and warn users about the limitations of such data, we determined that although this dataset does have potential for answering our research questions, it lacks the context present in the real NKR data.

²⁹ <https://iknl.nl/en/ncr/synthetic-dataset>

³⁰ <https://medium.com/towards-artificial-intelligence/analysis-of-a-synthetic-breast-cancer-dataset-in-r-1aba3cfe5a87>

NKR dataset

The process of obtaining data is two-stage. First, a submission form must be completed which is subsequently checked by IKNL for completeness. The initial IKNL submission form can be found [here](#) (in Dutch) and describes our objectives, research question, patient population, data flow, and analysis plan. The submission was internally discussed for completeness, relevance, and for conflict with another pending request. It is here that we were contacted by IKNL that our request had several overlapping parts with another submission. However, this was not a data request from an outside research body, but an IKNL study proposal for additional funding from ZonMw. By request of IKNL we had a Teams meeting with the main applicant of the proposal - Sabine Siesling. Their study proposal (which is in Dutch) was shared with us, and can be found [here](#). The idea was to discuss the length and magnitude of potential overlap. During the discussion, we were notified that our submission (NKR code: K23.185) would be placed in the earliest round of the NABON-BOOG committee. During our meeting with Siesling, we obtained valuable information regarding their application, and the process, and we also learned that Siesling herself would be on that committee.

IKNL judgement form

From this meeting, we received the final judgement form which was to be sent to NABON-BOOG for our meeting on the 18th of July 2023. Our final form can be found [here](#) and it contains the same information already sent in the initial submission form. The only notable exception is the list of variables that we would like to have access to, to complete our analysis plan.

Following our invited pitch during the meeting of the 18th of July³¹ we [received](#) a mail that same day in showing that our request would not be honored. The official document can be found [here](#). To qualify for a next round, additional questions would have to be answered which revolved around two key issues:

1. It was unclear to IKNL what we would have liked to achieve with our research and how this would benefit cancer patients.
2. Our list of supporting physicians (Dr. Sabine Pinedo and Dr. Roderik A. Kraaijenhagen) was deemed insufficient, considering they are not oncologists or radiologists, but an internist and a cardiologist.

Because of our stringent timeline, and our duty to ZonMw to choose the best possible method, we [suggested](#) teaming up. In summary, we offered Siesling to [receive](#) from us [our codes](#) and findings which could then be used for additional research. Our proposal was [declined](#) by IKNL for reasons of not wanting to have two similar studies. After a last final and stringent [request](#), which was [declined](#), we decided to discontinue our efforts to obtain the

³¹ Prof. Siesling was also part of that committee and was active during the meeting in asking questions about the feasibility of our proposal.

IKNL data which we believe would have offered us the best possible information. Considering our tight schedule, we did not believe that we would meet their requests. In addition, we also did not believe that adding more information or people would provide us with the necessary green light. We will describe the impact of all of this, and our views on the data landscape of the Netherlands, more extensively in the [discussion](#) part of this report.

Back to the original plan with more focus and four scenarios to explore

With the absence of patient-level data, we decided to return to the original plan of working with ecological data. Because of the time lost obtaining CBS and IKNL data, we decided to only use open data even if this meant losing valuable information and limited external validity. In addition, we decided to focus only on the second research question, but with a slightly altered approach.

First, we determined that we would focus our efforts on finding evidence for four potential scenarios (or hypotheses). These are:

1. There is no excess mortality in 2020 and 2021.
2. The excess mortality from 2020 and 2021 can be fully explained by Covid-19.
3. The excess mortality from 2020 and 2021 can (partly) be explained by the temporary cessation of population screening.
4. The excess mortality from 2020 and 2021 can (partly) be explained by delayed care of already diagnosed cancer patients.

Secondly, we conducted two rapid reviews: one on clinical prediction models for breast cancer, and one on the influence of delayed care during Covid-19 on breast, cervical and colon cancer mortality. Our reason for focusing on a clinical prediction model for breast cancer resulted from the availability of the synthetic breast cancer set. This is the only dataset available that has patient-level data. Despite its limitations, and possible mistakes, it would allow us to model and identify important predictors for survival. Guided by the outcomes of the rapid review, it would allow for a more targeted approach to investigating the open data sources. The rapid review on delayed care was additionally done to look for information that could support or discredit our findings in the open data.

Third, we added Jan Bonte, neurologist, as a third clinical expert. This would allow us to receive expert opinion on the plausibility of each of the four scenarios.

In summary, our efforts focused completely on our original second research question:

What is the effect of the Covid-19 pandemic on cancer screening in The Netherlands and how did this impact (excess) mortality in 2021?

We will describe our methods, and our [results](#), in full in the following two sections.

Methods

Although the Dutch government openly promotes the use of open data, there is no national data dictionary of all available Dutch databases. Also, it is not unthinkable that international databases have information on the Netherlands which are not directly available from the national databases. In addition, multiple agencies collect data on the same topic, such as [mortality](#), but use different methods to identify and collect. For instance, although one may expect that observed mortality follows the same definition regardless of database used, this cannot be said for a model-construct such as [excess mortality](#). Hence, we will try to be as precise as possible in describing each component of our study and the definitions used.

Components of our study and definitions used

To answer the first two questions, we already indicated that we would be using OpenDIS. However, OpenDIS is not the only relevant database in the Netherlands. For instance, VEKTIS has information on healthcare use as well. To further guide our search for relevant databases, we must first take a better look at the individual components of our study and find available definitions to investigate that component. For instance, to answer our overarching objective (*to estimate the effect of delayed cancer screening on cancer mortality during the pandemic for breast, cervical and colorectal cancer on (excess) mortality*) it is best to start with the concept of excess mortality.

[Excess mortality](#) is a model-construct which, in its simplest form, the difference between observed and expected mortality.³² The World Health Organization defines it as “*The mortality above what would be expected based on the non-crisis mortality rate in the population of interest*”. Although we must assume, for simplicity’s sake, that observed mortality is factual, excess mortality is not. To estimate excess mortality³³, the CBS must revert to modelling the expected mortality. The CBS has published some information on how they do this, but the information is high over.³⁴

To model expected mortality, you must also model [life expectancy](#). By modelling how long a specific Dutch citizen will live, you can estimate the expected death rate per year. From this expected death rate, you can deduce the observed death rate. These models are

³² <https://www.cbs.nl/nl-nl/nieuws/2023/05/eind-januari-geen-oversterfte-meer/oversterfte-en-verwachte-sterfte>

³³ <https://www.cbs.nl/en-gb/news/2023/28/excess-mortality-in-june>

³⁴ <https://www.cbs.nl/nl-nl/longread/rapportages/2022/sterfte-en-oversterfte-in-2020-en-2021/4-sterfte-en-doodsoorzaken-tijdens-de-covid-19-epidemie-in-nederland>

extrapolation models and a more detailed description of the prognostic model used for Dutch citizens³⁵ can be found [here](#). In short, these models are dynamic models that consider the year somebody is born, their gender and some health-indicators such as smoking. Hence, to get full insight into excess mortality, we must also show information about life expectancy, [mortality](#), and [cause of death](#).

The second part that needs to be explicitly described is [the Dutch screening system](#). By including [breast](#), [colon](#), and [cervical](#) cancer we have made it our objective to study three cancer types for which a nation-wide screening network has been set in place (i.e., population screening). Of the three, breast cancer screening has been around the longest. We will therefore show how the screening has developed over time, before connecting it to [life expectancy](#), [cause of death](#) and [delay of care](#). In addition, we will show the trajectories of the [incidence and prevalence](#) of the three cancer types across the years. Since we know that not everyone gets their first referral via the nation-wide screening process (for instance, a person could detect breast cancer themselves via self-examination), we will show the development of cancer prevalence next to the development of screening prevalence. Here, the years 2020 and 2021 are of course key.

Last, but not least, we must be very specific about the health-care data that we have used. The OpenDIS dataset is by far the greatest dataset containing information on DBC³⁶ (Diagnose-handelcombinatie or Diagnosist-Treatment Combination). Each DBC is a lengthy code which provides a health-professional or health-insurer about the diagnosis, treatments, and follow-ups a patient received. In summary, a DBC provides detailed information on the entire health-care trajectory of a patient, and thus forms the basis of the bills that need to be paid³⁷. In the absence of patient-level information, the DBC is the best possible method for investigating which type of diagnosis, treatment and follow-up was billed for across the years.

Below you can find the definitions used for each of the important outcomes of our study.

Outcome of interest	Definition used
Mortality	The total number of people pronounced dead by an authorized medical doctor in a given period. ³⁸
Mortality rate	The estimated total number of deaths in a population of a given sex and/or age, divided by the total number of this population,

³⁵ <https://www.cbs.nl/nl-nl/achtergrond/2013/26/bevolkingsprognose-2012-2060-model-en-veronderstellingen-betreffende-de-sterfte>

³⁶ <https://www.zorgwijzer.nl/faqdbc>

³⁷ <https://www.opendisdata.nl/msz/zorgproduct>

³⁸ <https://www.cbs.nl/en-gb/our-services/methods/definitions/mortality>

	expressed per 100,000 population, for a given year, in a given country, territory, or geographic area. ³⁹
Mortality risk	Risk to die in a given year. ⁴⁰
Life expectancy	Table showing how many of 100,000 newborn boys or girls will reach the age of 0.5, 1.5, 2.5 years etc. on the basis of the mortality rate during a given period or birth cohort. ⁴¹
Life expectancy	The number of years someone of a certain age still is expected to live, according to a life table. ⁴²
Life expectancy at birth	<p>The average number of years that a newborn could expect to live, if he or she were to pass through life exposed to the sex- and age-specific death rates prevailing at the time of his or her birth, for a specific year, in each country, territory, or geographic area.⁴³</p> <p>The number of years someone is expected to live at birth according to the life table.⁴⁴</p>
Excess Mortality	The mortality above what would be expected based on the non-crisis mortality rate in the population of interest. ⁴⁵
Cause of Death	A distinction is made between primary and secondary causes of death. The primary cause of death is the disease, situation or event that started the chain of events resulting in death. Consequences or complications of this are usually considered secondary causes of death, in the same way as other diseases present at the time of death that may have contributed to the death. ⁴⁶

Datasets used

For this study we used the following datasets which we will describe using the information on their respective websites. In the [Results](#) section, we will start by showcasing the descriptive content of each of the datasets used.

CBS

CBS's statutory task is to compile statistics on a wide range of social topics and to make the results publicly available.⁴⁷ To this end, CBS collects data from individuals and businesses. These collected data are processed into statistics. CBS uses various methods to collect data. Individuals and companies are surveyed; this is mainly done digitally, but sometimes also in

³⁹ <https://www.who.int/data/gho/indicator-metadata-registry/imr-details/1157>

⁴⁰ <https://www.cbs.nl/en-gb/our-services/methods/definitions/mortality-risk>

⁴¹ <https://www.cbs.nl/en-gb/our-services/methods/definitions/life-expectancy-table>

⁴² <https://www.cbs.nl/en-gb/our-services/methods/definitions/life-expectancy>

⁴³ <https://www.who.int/data/gho/indicator-metadata-registry/imr-details/3131#:~:text=Definition%3A,%2C%20territory%2C%20or%20geographic%20area.>

⁴⁴ <https://www.cbs.nl/en-gb/our-services/methods/definitions/life-expectancy-at-birth>

⁴⁵ <https://www.who.int/data/sets/global-excess-deaths-associated-with-covid-19-modelled-estimates>

⁴⁶ <https://www.cbs.nl/en-gb/our-services/methods/definitions/cause-of-death>

⁴⁷ <https://www.cbs.nl/en-gb/about-us>

writing or in person. In recent decades, CBS has increasingly used existing registrations. Examples are the population register or the files of the Chamber of Commerce. In addition to key registers from the government, CBS has also started to use registrations from companies, for example checkout data from supermarkets to calculate price trends. The big advantage of using registers is that CBS does not have to approach individuals and companies as often. This reduces the survey burden.

Eurostat

Eurostat is the statistical office of the European Union (EU), situated in Luxembourg. Its task is to provide the EU with statistics at a European level that enable comparisons between countries and regions.⁴⁸ Eurostat does not work alone⁴⁹ and is part of a larger European Statistical System (ESS) which is a partnership between Eurostat, and the National Statistical Institutes, and other national authorities present in each EU country. This partnership also includes the European Free Trade Association (EFTA) countries. For further details, there is a list⁵⁰ available showing the national statistical institutes and other national authorities designated by the countries to develop, produce, and disseminate European statistics.

IKNL

The Netherlands comprehensive cancer organization (IKNL) is the quality institute for oncological and palliative research and practice.⁵¹ IKNL collaborates with healthcare professionals and managers and patients on the continuous improvement of oncological and palliative care. The Netherlands Cancer Registry is the national registration since 1989, providing statistics on cancer in the Netherlands. It is the only oncological hospital registry in the Netherlands with data on all cancer patients including cancer diagnosis, tumor staging (according to the TNM-classification developed and maintained by the Union for International Cancer Control (UICC)), tumor site (topography) and morphology (histology) (according to the WHO International Classification of Diseases for Oncology (ICD-O-3)), co-morbidity at diagnosis and treatment received directly after diagnosis. The NCR is linked to the International Agency for Research on Cancer (IARC)⁵² and delivers anonymous data to the European database of the European Network of Cancer Registries (ENCR).⁵³

OECD

The Organisation for Economic Co-operation and Development is an intergovernmental organisation with 38 member countries, founded in 1961 to stimulate economic progress

⁴⁸ https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Main_Page

⁴⁹ <https://ec.europa.eu/eurostat/web/european-statistical-system/overview>

⁵⁰

https://ec.europa.eu/eurostat/documents/13019146/13574152/20230803_List_other_national_statisticalAuthorities_PL.pdf/c199058f-eb0d-7cc3-ddcf-990878b40831?t=1691048761794

⁵¹ <https://iknl.nl/en/about-iknl>

⁵² https://en.wikipedia.org/wiki/International_Agency_for_Research_on_Cancer

⁵³ <https://www.enrcr.eu/>

and world trade.⁵⁴ It is a forum whose member countries describe themselves as committed to democracy and the market economy, providing a platform to compare policy experiences, seek answers to common problems, identify good practices, and coordinate domestic and international policies of its members. For each of the countries included there is a country statistical profile available in the OECDiLibrary.⁵⁵ In addition, there are numerous databases freely available based on theme, country, or education. For the Netherlands, relevant data can be captured with regards to life expectancy at birth⁵⁶, deaths from cancer⁵⁷ and health care utilization.⁵⁸

OpenDIS

The DBC information system (DIS) of the NZa (Nederlandse Zorgautoriteit)⁵⁹ receives and manages information on all treatment processes in medical specialist, mental health, forensic and rehabilitation care.⁶⁰ Healthcare providers in these sectors have a legal duty to supply this data. Via OpenDIS the NZa wants to provide insight into the medical-specialist care provided in the Netherlands. This website only shows generic data on medical-specialist care and displays data that falls under the Open Government Act (WOO). Information that can be traced back to individual patients and institutions is not accessible. In summary, you can see how many patients received certain diagnoses, the corresponding care products and care activities and the average selling price. The information structure follows the structure of the DBC system, in which a specialty opens a DBC track under a certain diagnosis and registers a DBC track within this. Within this pathway, care activities are registered. This registration then leads to a certain DBC care product. The OpenDIS website only contains care activities with a profile class (ZPK; Zorgprofielklasse) of 1-8, 14 or 16. The table below shows a translated description⁶¹ of each of the profile classes included.⁶²

ZPK	Description
1	Outpatient visit, primary care visit and remote consultation
2	Day care
3	Clinic
4	Diagnostic activities
5	Surgical operations

⁵⁴ <https://en.wikipedia.org/wiki/OECD>

⁵⁵ https://www.oecd-ilibrary.org/economics/country-statistical-profiles-key-tables-from-oecd_20752288

⁵⁶ <https://data.oecd.org/healthstat/life-expectancy-at-birth.htm>

⁵⁷ <https://data.oecd.org/healthstat/deaths-from-cancer.htm>

⁵⁸ https://www.oecd-ilibrary.org/social-issues-migration-health/data/oecd-health-statistics/oecd-health-data-health-care-utilisation_data-00542-en

⁵⁹ <https://www.nza.nl/>

⁶⁰ <https://www.opendisdata.nl/dis/over>

⁶¹ <https://www.deepl.com/en/translator>

⁶²

https://demedischspecialist.nl/sites/default/files/Registratiewijzer_%20Federatie_Medisch_Specialisten_2021.pdf

6	Other therapeutic operations
7	Imaging diagnostics
8	Clinical chemistry and hematology
14	Rehabilitation
16	Geriatric rehabilitation

The average national (historical) prices are calculated based on all declared reimbursements supplied from 2012 onwards. This includes both contracted and non-contracted care. The reimbursements include hospital costs as well as medical specialists' fees. Information on current local prices is available from health insurers. No average is calculated if fewer than five healthcare providers have registered for the healthcare product, or fewer than 100 in total are registered. DBC care products with no selling price are also not included. To estimate the average, lowest and highest 10% of prices are excluded. A description of the entire content of the database can be found [here](#).⁶³

Healthcare products (DBC) of interest (OpenDIS)	
DBC code	DBC description (Dutch)
028899046	Kijkoperatie van de dikke darm bij de screening op kanker van dikke darm of endeldarm
028899048	1 of 2 polikliniekbezoeken/ consultaties op afstand bij de screening op kanker van dikke darm of endeldarm
219899014	Preventieve operatie bij de screening op erfelijke vormen van kanker
020107002 – 020107057	Behandeling borstkanker
020108044 – 020108090; 020108233 – 020108236	Behandeling baarmoederhalskanker
029199003 – 029199106; 029199269 – 029199272; 029199286	Behandeling dikke darmkanker

Our World in Data

Our World in Data (OWID) is a scientific online publication that focuses on using numbers and statistics to describe the current state of the world with regards to large problems such as poverty, disease, hunger, climate change, war, existential risks, and inequality.⁶⁴ Data is freely available on many topics for many of the countries in the world. For this study, we have used data with regards to cancer mortality⁶⁵, excess mortality⁶⁶, and causes of death.⁶⁷

⁶³ https://www.opendisdata.nl/download/DIS00981_Beschrijving_downloads_Open_DIS_data_1.0.pdf

⁶⁴ https://en.wikipedia.org/wiki/Our_World_in_Data

⁶⁵ <https://ourworldindata.org/cancer>

⁶⁶ <https://ourworldindata.org/excess-mortality-covid>

⁶⁷ <https://ourworldindata.org/causes-of-death>

State of Public Health and Health Care

The State of Public Health and Health Care (de Staat van Volksgezondheid en Zorg; staatvenz) provides up-to-date to support policy monitoring and accounting, and is built specifically for the Ministry of Health and the Members of the House of Representatives.⁶⁸ The website is a product of collaboration in the form of a consortium. Collaborating parties are the National Institute for Public Health and the Environment (RIVM)⁶⁹, Statistics Netherlands (CBS), The Dutch Healthcare Authority (NZa), The Dutch National Health Care Institute (ZIN), The Netherlands Institute for Health Services Research (NIVEL), The Netherlands Institute for Social Research (SCP), The Trimbos Institute for mental health, and the Netherlands Youth Institute (NJI). We have used this website to gather data on incidence and prevalence of breast, cervical and colon cancer, and its related screening procedures.

VEKTIS

VEKTIS⁷⁰ is a Dutch company in the health care sector which collects geodata and healthcare data. VEKTIS organises data of insurance expense reports and does independent research for parties in healthcare.⁷¹ VEKTIS delivers service and products mainly to insurers in healthcare, healthcare providers, (local) governments and software deliverers. VEKTIS publishes reports on healthcare related topics, such as waiting times and numbers of healthcare workers per postal code area. In addition, it harbors the AGB-registry. An AGB is a unique 8-digit code issued by healthcare providers for which the first two digits indicate the type of provider.⁷² From VEKTIS we have used data on the type of health care provided for the years 2011 until 2021 on a postal level containing three digits of the postal code (i.e., postcode3). This translated to information from around 500 regions in the Netherlands.

VZinfo

The website www.vzinfo.nl contains information about Dutch public health and care, and harbors numbers s meant for Dutch professionals who are active in the field of public health. As such, it is the gateway to information about health and disease, risk factors, care, and prevention in the Netherlands⁷³. VZinfo.nl is developed and coordinated at the National Institute of Public Health and the Environment (RIVM), and contains efforts from experts, research-institutes, and universities. The ministry of Health, Welfare and Sports (VWS) has commissioned the website. On the website there are multiple themes such as mortality, cost of illness and healthcare. For this study, we used the website to analyze data on the mortality associated with the three types of cancer.

⁶⁸ <https://www.staatvenz.nl/english>

⁶⁹ <https://www.rivm.nl/en>

⁷⁰ <https://www.vektis.nl/>

⁷¹ https://data.europa.eu/sites/default/files/use-cases/netherlands_-_vektis.pdf

⁷² <https://nl.wikipedia.org/wiki/AGB-code>

⁷³ <https://www.vzinfo.nl/english>

Human Mortality Database

The Human Mortality Database (HMD)⁷⁴ is the self-proclaimed world's leading scientific data resource on mortality in developed countries. The HMD provides detailed high-quality harmonized mortality and population estimates to researchers, students, journalists, policy analysts, and others interested in the human longevity. The HMD also claims to follow open data principles but has recently changed its user agreement for accessing its data source. Since June 15th of 2022, data requests can only be made if a requester has registered. Fortunately, there are two large GitHub pages containing HMD data which we can freely access. The first is maintained by [Ariel Karlinksy](#) and the second by [Dmitry Kobak](#) who have published on their work on excess mortality during Covid-19 using the HMD data,^{75 76 77} and who work is one of the primary database to be used by Our World in Data (OWID).⁷⁸ Hence, for this study, we used the data from both sources on excess mortality.

Building a clinical prediction model

The statistical analysis of this study was focused on building a clinical prediction model for breast cancer survival. Breast cancer is the largest population of the three so any effect through delayed care would have an immediate impact on the excess mortality numbers. Since we did not obtain real IKNL data for breast cancer, colon cancer, or cervical cancer, we used the synthetic breast cancer dataset we did have access to. However, this dataset has less variables included. For instance, there is no information on when a particular treatment was provided, so we cannot estimate "time-since" variables which would provide us with useful insights on treatments trajectories. Furthermore, we already mentioned that the descriptive nature of the data is limited, since that is the key to synthetic data. However, if constructed properly, it should provide us with the ability to detect any "real" relationships present between the variables.

Building a clinical prediction is not an easy task and should be handled with care when used for clinical decision making.⁷⁹ Several studies have indicated that the current set of prediction models supporting breast cancer decision making are at a high risk of bias,⁸⁰ but this is almost inevitable as models are fully dependent on the data obtained and humans (and healthcare) are more complex than algorithms. Models can easily paint alternative realities of situations that do not exist,⁸¹ classifying humans into combinations of patient, clinical and treatment characteristics.

⁷⁴ <https://www.mortality.org/>

⁷⁵ <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7852240/>

⁷⁶ <https://www.who.int/publications/item/estimating-global-and-country-specific-excess-mortality-during-the-covid-19-pandemic>

⁷⁷ <https://doi.org/10.26633/RPSP.2022.19>

⁷⁸ <https://ourworldindata.org/grapher/excess-mortality-p-scores-projected-baseline>

⁷⁹ <https://doi.org/10.1016/j.breast.2023.01.006>

⁸⁰ <https://doi.org/10.1016/j.jclinepi.2022.10.016>

⁸¹ <https://www.walburgers.nl/nl/book/9789462499928/van-aardbeving-tot-zooenoze>

Therefore, our approach to building a clinical prediction model, and using it to gain insight into the survival probability of breast cancer patients, was multimodal:

1. First, we conducted a rapid review on available clinical prediction models for breast cancer with a special focus on the predictors deemed important. This way, our clinical prediction model is not just based on the covariance matrix of synthetic data.
2. Second, we will extract the methods and findings from [our previous work](#). Published on November 22 in 2021, the analysis described in a Medium post shows multiple ways of trying to create a clinical prediction model for breast cancer.
3. We will combine our methods used then with the predictors identified in the literature to build a new model.
4. Last, but not least, we will identify the clinical characteristics of the ‘average’ breast cancer patient and include them in the model we created, and the models freely available to obtain survival curves. Our focus will be on the first two years following the diagnosis.

Rapid review of clinical prediction models for breast cancer.

To guide our second, and most important, research question (*What is the effect of the Covid-19 pandemic on cancer screening in The Netherlands and how did this impact (excess) mortality in 2021?*) we will build a clinical prediction model for breast cancer using the synthetic dataset. The building of such a model needs to be driven by clinical input that is available in literature, and the data is made available via IKNL. Despite various limitations we do believe that it is worth a try to build such a model. The aim of a clinical prediction model for breast cancer is multifold:

1. To identify clinical predictors that may influence the survival of breast cancer patients.
2. To estimate how many patients should have died and did die.
3. To estimate how many patients are likely to die from breast cancer in 2020, 2021 and 2022.
4. To assess the validity of such a model via literature and internal validity procedures (to be discussed in the statistical analysis part).
5. To estimate the possibility that the excess mortality seen in 2020, 2021 and 2022 is due to difference in diagnosis and care received.
6. To estimate possible excess mortality in the years to come from reduced diagnosis and care in 2020.

We already mentioned that building a clinical prediction model on data alone is not sufficient. For years, researchers and clinicians have tried (and are still trying) to determine who will die and when from breast cancer. This is a very difficult exercise to do, and so we cannot just rely on a single data set alone and must also look to the findings of others. One

of the first models we will build will be based entirely on the findings of our literature search. The rest will be either completely data-driven, or a hybrid, which we will explain later in the Statistical Analysis section.

Given the limited time and availability of resources, we will not conduct a systematic review following PRISMA guidelines.⁸² Instead, we will use Google Scholar to search for systematic reviews and meta-analyses on which markers are most informative in predicting the survival of breast cancer patients. Methodological assessment will not follow the Cochrane Handbook for Risk of Bias,⁸³ nor will we use any other checklist. Instead, we will assess as best as possible how often several markers are labeled as important, and we will extract the coefficients of each marker if the model and publication allow this. Machine learning models especially are not able to provide direct strengths of evidence in terms of coefficients, but rather in terms of variable importance (which is often normalized from 0 to 100%).

In the end, the aim of this literature search is to get a list of important variables from various sources which we can then use to build our first clinical prediction model. A full systematic review and meta-analysis based on all available breast cancer models would require a separate endeavor. We hope that this has already been done.

Articles will be read and reviewed by one person (Marc Jacobs) and by an additional person (Ronald Meester) if necessary. Full-text articles will be extracted from the website of the publisher, or in case of a paywall, from SciHub.⁸⁴

Because of the inherent limitations of our data, we also conducted a rapid review on the possible influence of delayed care during Covid-19 on breast, cervical and colon cancer mortality.

Analysis plan of our previous work

In our medium post, we described the following parts in the form of a blog post unrelated to this research, but with the exact same aim. Hence, our analysis plan looked like this:

1. Full-scale graphical exploration of distributions of variables, missing variables, and relationships.
2. Classifying each population into clinically meaningful groups based on underlying morphology and localization. For continuous variables, we determined the optimal cut point for one or multiple continuous variables at once, using the maximally selected rank statistic.⁸⁵
3. Designing a risk classification model where the probability of death is a function of time, and the variables included.⁸⁶

⁸² <http://www.prisma-statement.org/>

⁸³ <https://methods.cochrane.org/bias/>

⁸⁴ <https://sci-hub.se/>

⁸⁵ <https://cran.r-project.org/web/packages/maxstat/vignettes/maxstat.pdf>

⁸⁶ <https://towardsdatascience.com/introduction-to-survival-analysis-in-cows-using-r-28b82c2821fb>

- a. First, we built numerous univariate survival models for right-censored data⁸⁷ using the Kaplan-Meier method.⁸⁸ This was done on almost all variables included.
 - b. Second, we started building multivariable models including variables deemed important. The way of building the models back then was based on what we deemed clinically important, but necessarily on literature. For this part, we fitted proportional hazards regression models via Cox.⁸⁹
 - c. For each of the models created, we extracted concordance statistics,⁹⁰ assessed coefficient estimates and standard errors, and looked at variance inflation factors⁹¹. All of this was done to assess the stability of the models and detect issues of multicollinearity. Cox regression models were further assessed via diagnostics of deviance, Schoenfeld, and Martingale residuals. We also assessed the information provided by continuous variables, like age, positive lymph nodes, and tumor size which were initially fitted using natural splines.⁹²
 - d. Fourth, we applied L1 (LASSO) and L2 (Ridge) regularization methods to detect if we could further downsize the models^{93 94}. When trying to create a clinical prediction model, any model, obtaining a parsimonious model is key⁹⁵. Therefore, we combined the L1 and L2 methods into an Elastic Net⁹⁶ from which we tried to delete or shrink coefficients. We used k fold-cross-validation to determine the optimal bias to be included in the models⁹⁷.
4. Validation of any model is important, but it is also very difficult. For instance, the model we built cannot be externally validated since we do not have patient-level data on the years of interest (2020-2021). However, we do have a lot of data which we can internally validate using bootstrapping which is what we did back then⁹⁸. Using 500 bootstrapped⁹⁹ samples, we created calibration methods for 1000, 2000, and 3000 days splitting the data in groups of 150 patients.¹⁰⁰ This way, we can see if the Cox regression models performed equally well in time.

⁸⁷ <https://blog.devgenius.io/survival-analysis-in-sas-kaplan-meier-cox-regression-time-varying-predictors-recurrent-events-4ae7cd95f8c0>

⁸⁸ <https://www.rdocumentation.org/packages/survival/versions/2.11-4/topics/survfit>

⁸⁹ <https://rdrr.io/cran/survival/man/coxph.html>

⁹⁰ <https://cran.r-project.org/web/packages/survival/vignettes/concordance.pdf>

⁹¹ <https://rdrr.io/cran/car/man/vif.html>

⁹² <https://www.rdocumentation.org/packages/survminer/versions/0.4.9/topics/ggcoxfunctorial>

⁹³ <https://cran.r-project.org/web/packages/penalized/index.html>

⁹⁴ <https://cran.r-project.org/web/packages/glmnet/index.html>

⁹⁵ <https://blog.devgenius.io/predictive-regression-using-splines-partial-least-squares-penalization-cross-validation-and-339b74a7e108>

⁹⁶ https://en.wikipedia.org/wiki/Elastic_net_regularization

⁹⁷ [https://en.wikipedia.org/wiki/Cross-validation_\(statistics\)](https://en.wikipedia.org/wiki/Cross-validation_(statistics))

⁹⁸ <https://cran.r-project.org/web/packages/rms/index.html>

⁹⁹ [https://en.wikipedia.org/wiki/Bootstrapping_\(statistics\)](https://en.wikipedia.org/wiki/Bootstrapping_(statistics))

¹⁰⁰ <https://rdrr.io/cran/rms/man/calibrate.html>

5. Besides building Cox regression models, which is the de facto standard for building multivariable semi-parametric models, we also tried a random forest approach to survival probability.^{101 102} These models are more difficult to decipher and do not offer coefficients but variable importance plots and tables. We assessed the models graphically by tracking the error rate across the number of trees, and by looking at some trees specifically. From the models we also extracted partial dependence plots showing the minimal depth of the variables included. Minimal depth for a variable in a tree equals to the depth of the node which splits on that variable and is the closest to the root of the tree. If it is low, then a lot of observations are divided into groups based on this variable.
6. In the end, we selected the best model, based on the Cox regression and random forest approaches and used that model to build a dynamic nomogram from which to extract survival probabilities based on patient, clinical, and treatment characteristics.

Analysis plan of our current work

Our current work calls for the integration of the methods used in our previous work and our findings from the rapid review on clinical prediction models for breast cancer. This integration will look like this:

1. We will extract important predictors from our previous work.
2. We will extract important predictors from our rapid review.
3. We will include all predictors in a Cox regression model.
4. We will assess the Cox regression model in the same way we assessed them in our previous work.
5. We will identify a final Cox regression model, build a dynamic nomogram, and extract survival probabilities.

What can we say at the end of all of this?

In the end, the risk classification model from step five serves as a basis for determining the following:

- a. To identify clinical predictors that may influence the survival of breast cancer patients.
- b. To identify characteristics of the “average” breast cancer patient.
- c. To identify characteristics of patients identified via screening.
- d. To estimate the survival probability of these patients.
- e. To assess the possibility that the excess mortality seen in 2020, 2021 and 2022 is due to difference in diagnosis and care received.

¹⁰¹ <https://cran.r-project.org/web/packages/ranger/ranger.pdf>

¹⁰² <https://cran.r-project.org/web/packages/randomForestSRC/index.html>

By far the most difficult step will be to “*assess the validity of the model based on external validity procedures.*” Let us explain more in depth, since is the difficulty of each model that is used for estimating “excess”. Our aim is to determine which predictor is most important for survival. Should the model indicate that TNM stage is very important, and that survival drops significantly from stage I to II we are able to use that information to look for a potential route towards excess mortality via reduced diagnosis and care. This is because we can check how many different TNM stages are detected each year. Should 2020 detect less TNM Stage II than usual, and we know that TNM Stage II has a limited survival rate, then we could say that it is not without a doubt that limited diagnosis did not influence survival because of reduced care. Of course, we want to look for markers related to specific care as well. Combining markers from patient-level clinical predictions to ecological data is circumstantial at best, but it is useful to inform.

However, the biggest limitation is that we still cannot externally validate our model. For instance, if our model predicts that, based on the current 2010-2019 patient population, 435 patients should have died from breast cancer and in fact 367 patients died, we do not know what exactly went wrong – if anything went wrong at all. The most honest answer is to say that it is almost impossible to disentangle excess mortality from limited external validity.

A noteworthy comment we want to make is that our models are built to predict survival, not mortality. Although survival and mortality models are of course related, and can be built using similar algorithms, survival models explicitly include time. A survival model estimates the risk of dying at a given time whereas mortality estimates death rates. If we were to build a mortality model, we would essentially estimate the risk of dying, but not the time of possible death. Because we want to also look beyond 2019, and we know that breast cancer has a relatively good survival rate, we want to build a survival model that will enable us to look beyond our data. Such a model can of course also be used to estimate mortality for a given year.

To further guide our assessments, we also conducted a rapid review on the possible influence of delayed care during Covid-19 on breast, cervical and colon cancer mortality.

Rapid review on the possible influence of delayed care during Covid-19 on breast, cervical and colon cancer mortality.

Like the literature search on clinical predictions models we will conduct a second literature search on the (potential) influence of delayed care. Because most of our data is ecological, and our clinical prediction model will be based on non-validated synthetic data, we believe it to be important to also get a view on what the literature has to say on delayed care and cancer mortality. Like the method above, this review is rapid and will not contain formal methodological assessment. Our aim is to search for possible routes that we may or may not confirm in our ecological data. It is here that the opinion of our clinical experts becomes important.

Clinical expertise

For each of the four scenarios drafted, we asked our clinical experts to what extent they believed that a scenario was plausible. All three have personally invested heavily in researching Covid-19, and both Sabine Pinedo and Roderik Kraaijenhagen are specialists in preventive medicine.¹⁰³ In the end, we sent them a [document](#) in which we outlined a summary of our findings and asked them to provide their vision of the four possible scenarios.

Results

As already mentioned in the methods section, we will first start by describing what we see from the open data that is available to us. We will use as many figures as possible to enable the reader to evaluate our descriptions. In addition, the reader becomes more familiar with the datasets with them. Then we will move forward to the prediction models using our statistical models and already developed models that are available. In the end, we will discuss the rapid review on delayed care and present the findings we received from the clinical experts. All of this will form the foundation for the discussion.

The descriptives are arranged by the various categories that we assessed:

1. [Life expectancy](#)
2. [Mortality](#)
3. [Excess mortality](#)
4. [Cause of death](#)
5. [Cancer incidence & prevalence](#)
6. [Cancer screening](#)
7. [Cancer survival](#)
8. [Health care expenditure](#)

Where possible, or required, we will combine data from the various categories since most categories have multiple data sources. The data source must be excluded as a possible explanation for whatever modelling we will conduct in the future.

For this part, we have tried to be as descriptive as possible, writing down what we saw from the graphics that we made based on the data that is openly available. High-resolution screenshots have been made available by clicking on the image in this document. You will automatically be directed to the [GitHub page](#) that is made available to all.

¹⁰³ <https://vital10.nl/>

Life expectancy

The CBS calculates every year what the life expectancy is for a person born in a specific year, by sex. In addition, they also present data on the life expectancy of someone given their current age and when they were born. For instance, one of the authors is born on the 11th of October in 1985, and male. His age on the 31st of December 2023 would make him 38. Combined, this will give him approximately another 40 years, meaning that his life expectancy (give or take) is around 78 years (Figure 1).

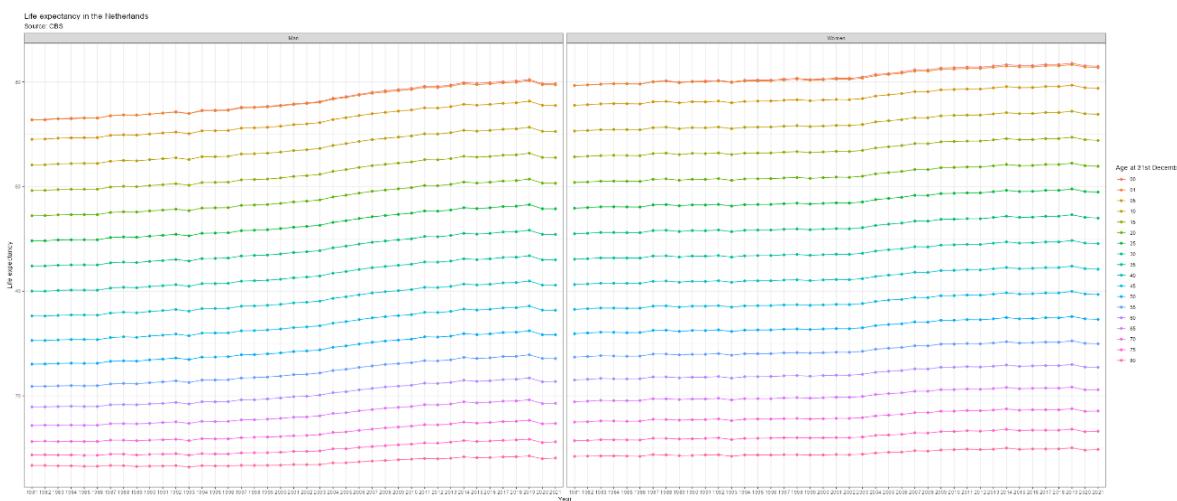


Figure 1. Life expectancy in The Netherlands (in years) per sex, year, and age at 31st of December.

Figure 2 shows the life expectancy for a male or a female when born in a specific year. From this graph, the author would have a life expectancy of around 72, give or take.

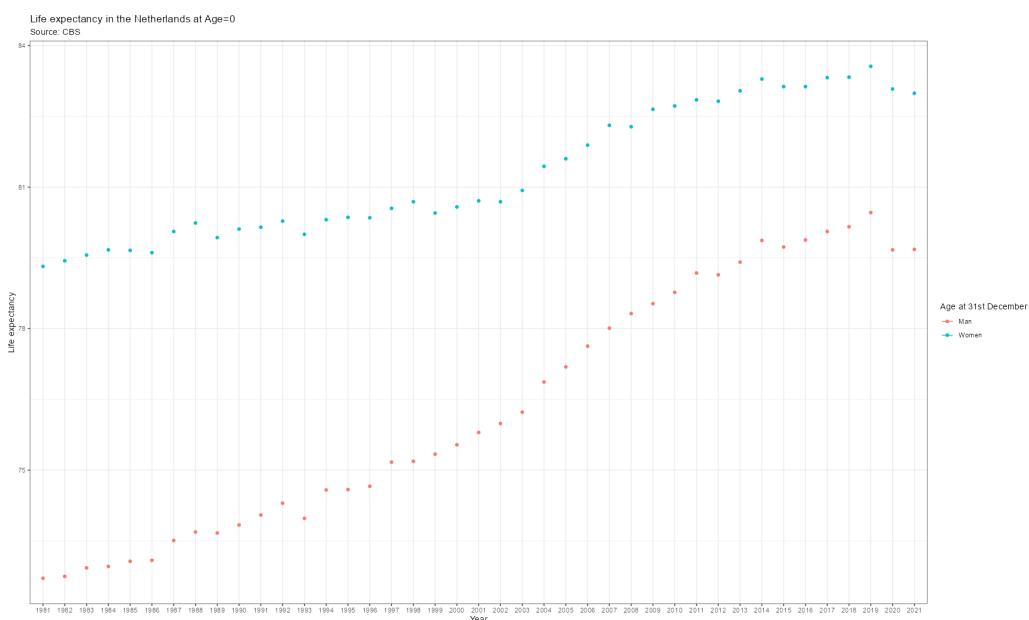


Figure 2. Life expectancy in The Netherlands (in years) per sex, year, and at age=0 at 31st of December.

Both graphs clearly show that females have a higher on average life expectancy than males. The graphs also show that the life expectancy of both groups has shown an increase across the years. However, in 2020 and 2021, the life expectancy in both groups has dropped dramatically and we assume that this is the result of Covid-19. We are not sure why exactly a newborn would have a life expectancy equal to that of 2013, because of Covid-19, when they are expected to die at the end of the century. By explicitly decreasing their life expectancy by several years, it seems that CBS is assuming that Covid-19 will be important for future generations to come.

Mortality

There are several sources that track the mortality of a country and we looked at all of them, combining their results where necessary to verify the validity of the data. Figure 3 below shows the all-cause mortality for the Netherlands which is extremely cyclical in nature. The biggest spike if of course the March-2020 which marked the beginning of the Covid-19 pandemic in the Netherlands.

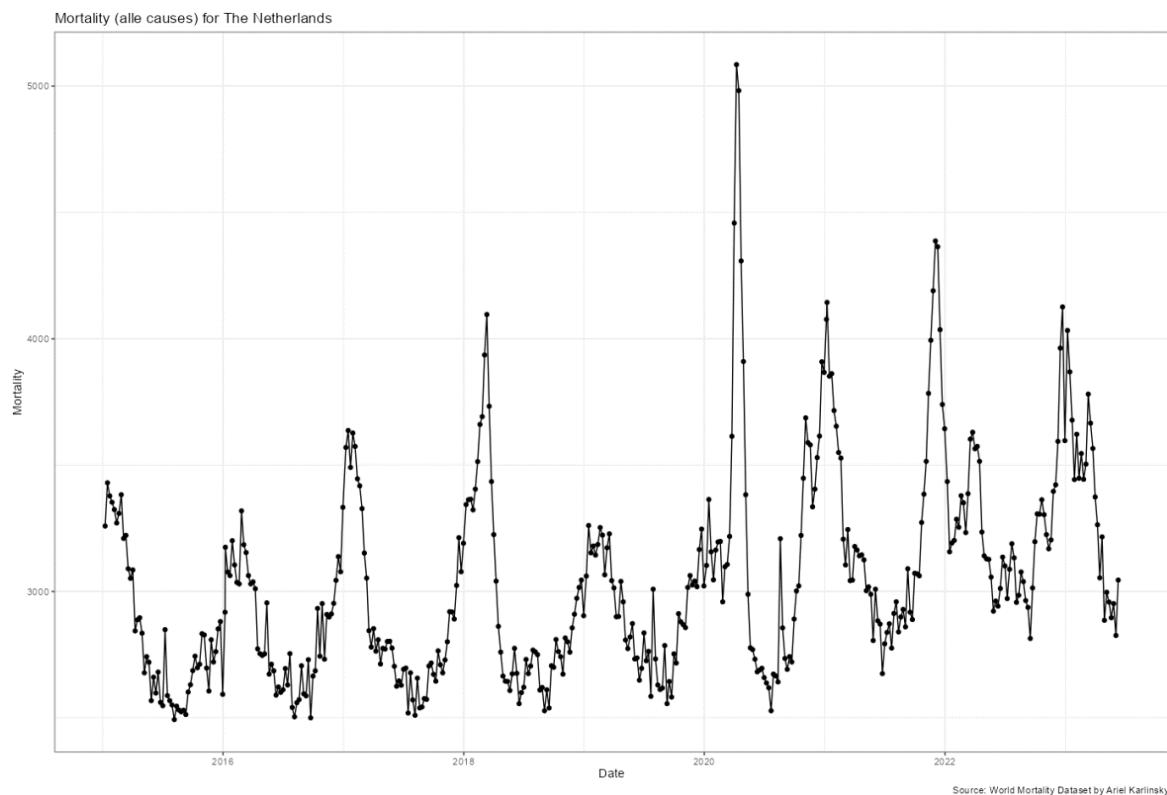


Figure 3. All-cause mortality for the Netherlands based on the World Mortality Dataset. Clear to see are the periodic spikes, with the biggest spike in March 2020 during the first Covid-19 wave. Spikes can still be seen all the way to 2023.

The mortality data are consistent across sources (Figure 4), but clearly show a trend reversal for 2020. Also, the slope of the mortality has increased somewhat, which can be better shown when plotting each month by using a heatmap (Figure 5). Clearly shown is the 2020 spike which is higher than anything seen in the past eight years. In addition, the winter

months show a clear increase in mortality which persisted well beyond 2022. Below the heatmap is a graph (Figure 6) depicting the curve for each year separately, highlighting only the last four years (2020-2023).

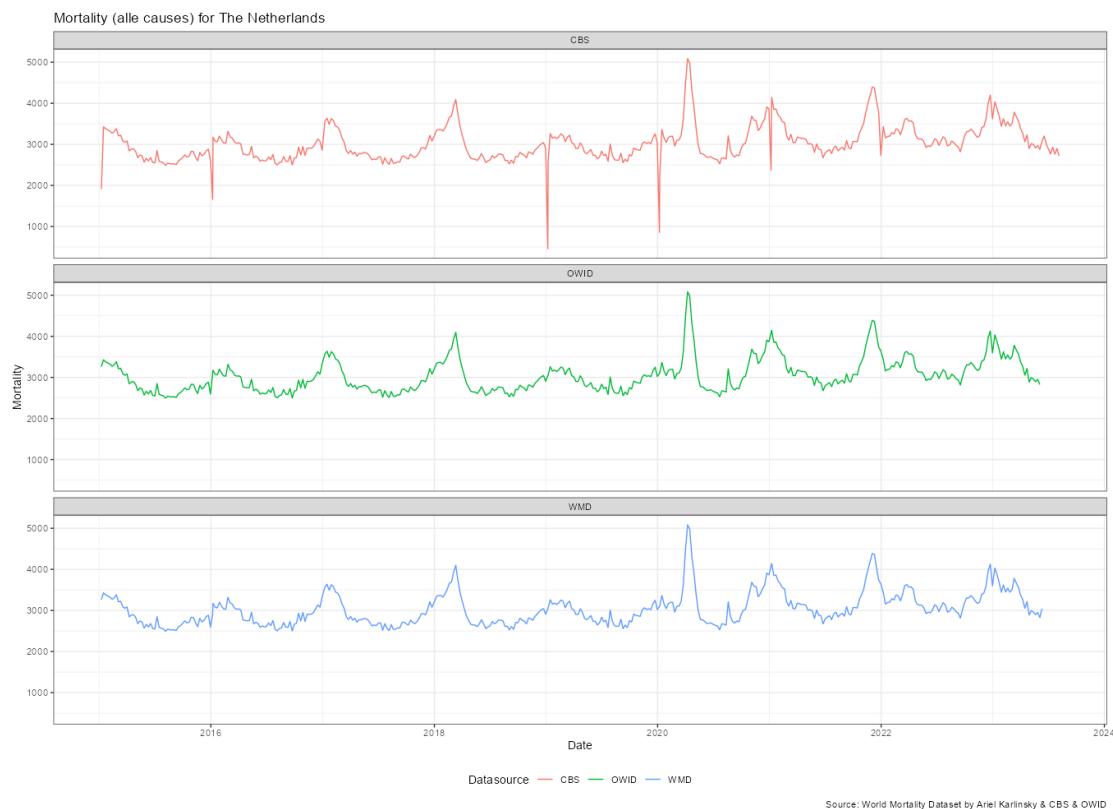


Figure 4. Checking the all-cause mortality data across sources.

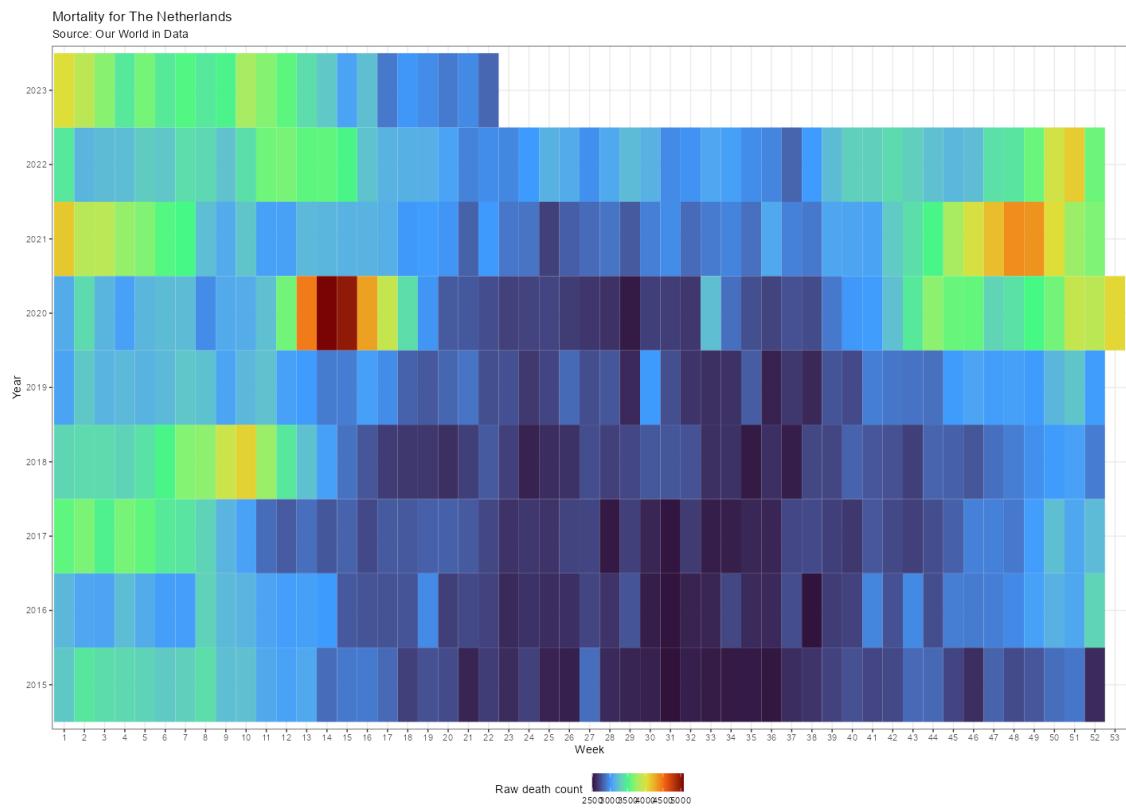


Figure 5. Heatmap of mortality data across months and years.

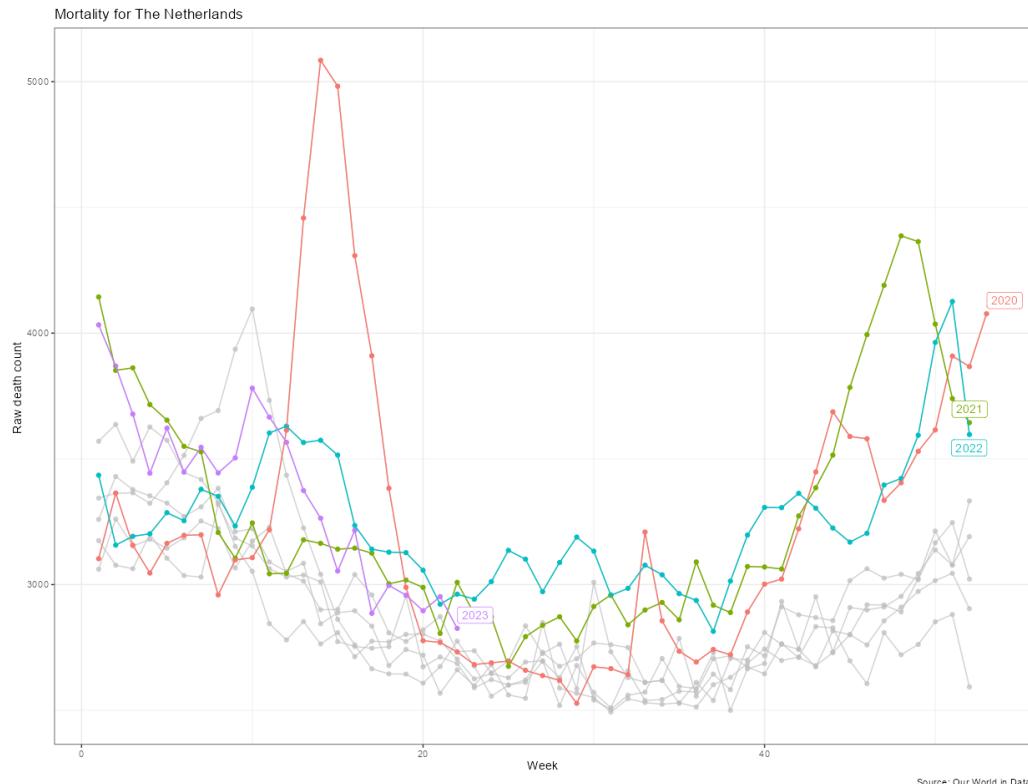


Figure 6. Mortality in the Netherlands, using the OWID data, showing mortality for several years. We highlighted the corona years (from 2020 onwards), showing the heavy spike in March 2020 and the consecutive spikes from 2020 to 2022.

From the life expectancy figures, we already saw that mortality is a function of age, so it makes sense to split the data further by age (Figure 7). Doing so makes it clear that the mortality has been on the rise for years, especially for the 80 years and older group. When looking at that group, the 2020 spike stands out, but only second to 2018.

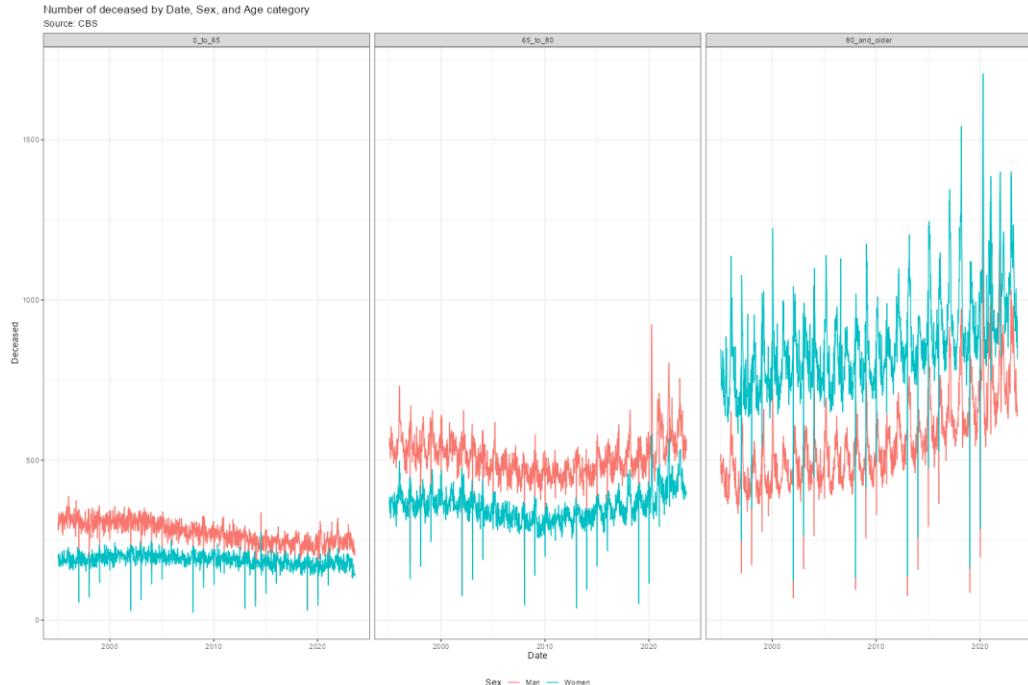


Figure 7. Mortality, using CBS data, per age category and across time. The strange correction factor in the CBS data is what provides the downward spikes.

The graph below (Figure 8) is basically the same graph as above but depicted differently. Once again, it is clear to see that mortality is a function of age and sex. Except for 2019, the mortality has been increasing for years.

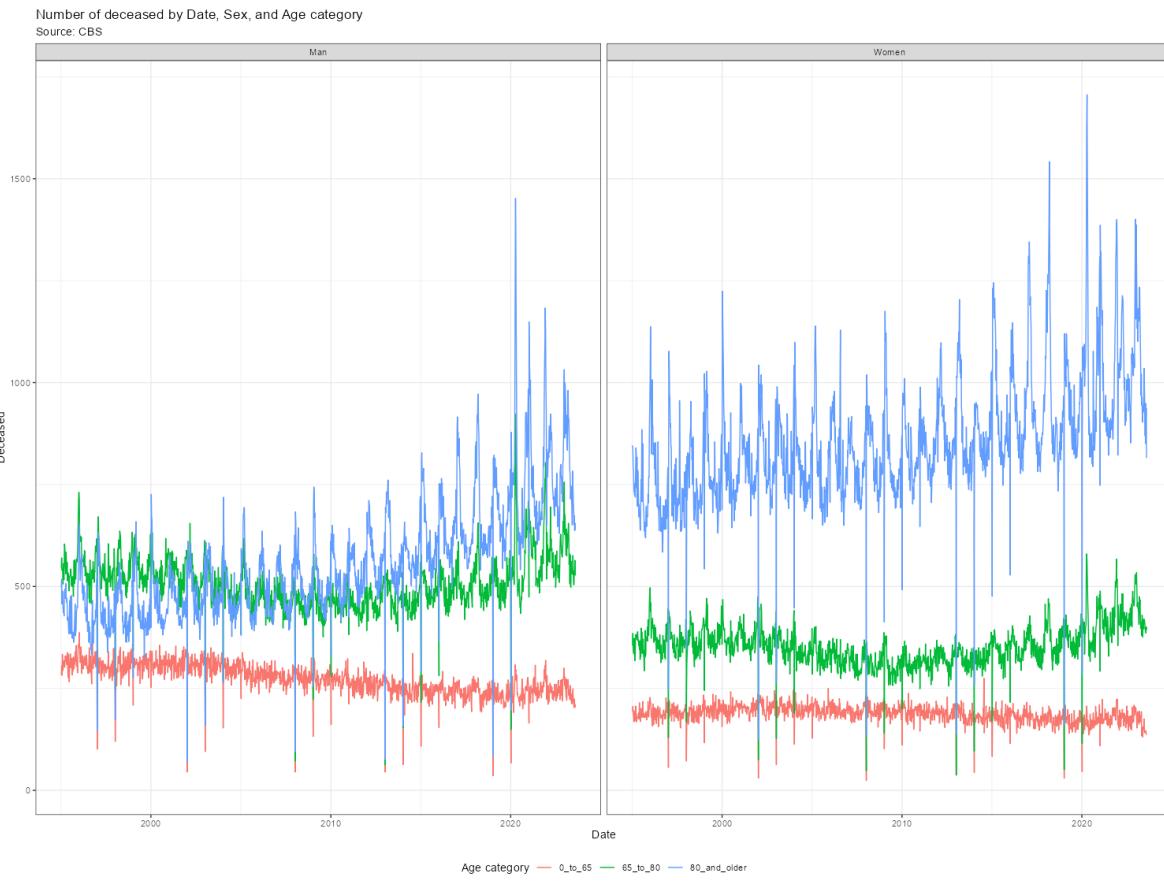
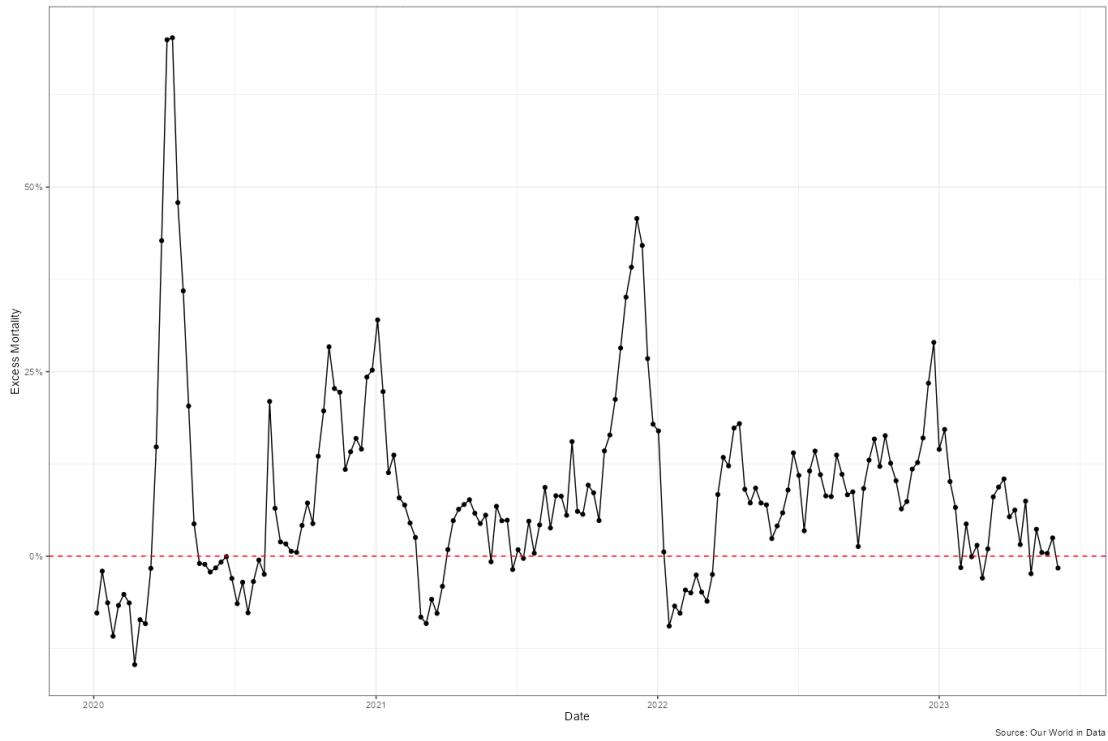


Figure 8. Mortality, using CBS data, per age category and sex. It has the same data as the previous figure.

Excess mortality

Combining life expectancy and mortality brings us to excess mortality. Although the direct model used is not clear to use (this is also true for life expectancy) it would make sense to include sex and age. The excess mortality metric is perhaps the most important metric in this document. This is because the excess death is a direct function of predicted death vs observed death, and any prolonged deviation is deemed worrisome. Also here, there are several sources that we used to fully understand the level of excess mortality. Below you can see the excess mortality as the percentage difference between the expected mortality and the observed mortality (Figure 9).

Excess Mortality (all causes) compared to projection for The Netherlands
The percentage difference between the reported number of weekly or monthly deaths in 2020–2023 and the projected number of deaths for the same period based on previous years.



Source: Our World in Data

Figure 9. This picture shows the excess mortality, using the OWID data, for the Netherlands.

Unfortunately, most sources only show excess mortality metrics from 2020 onwards, which makes it very difficult to get a wider picture of the potential role of Covid-19 and delayed care. Clear to see from the picture above are the large spikes in March 2020 and the end of 2021 and 2022. From March 2020 onwards there have been very little months in which there is a ‘deficit’ in deaths.

The next graph (Figure 10) shows how the excess mortality is derived as a function of the expected and observed mortality. Since we do not know exactly how the expected mortality is calculated we are left with assessing the difference, which is the excess. However, what we can see very clearly is that the expected mortality is cyclic and steady.

Figure 11 shows both all-cause mortality and excess mortality including the 95% confidence bands. Very clear to see is the stable cyclical pattern used to determine the expected mortality, and the underlying much more noise all-cause mortality. In essence, excess mortality should only be called ‘excess’ if the observed all-cause mortality exceeds the upper band. However, it seems that the average expected mortality is used to determine the number of excess deaths.

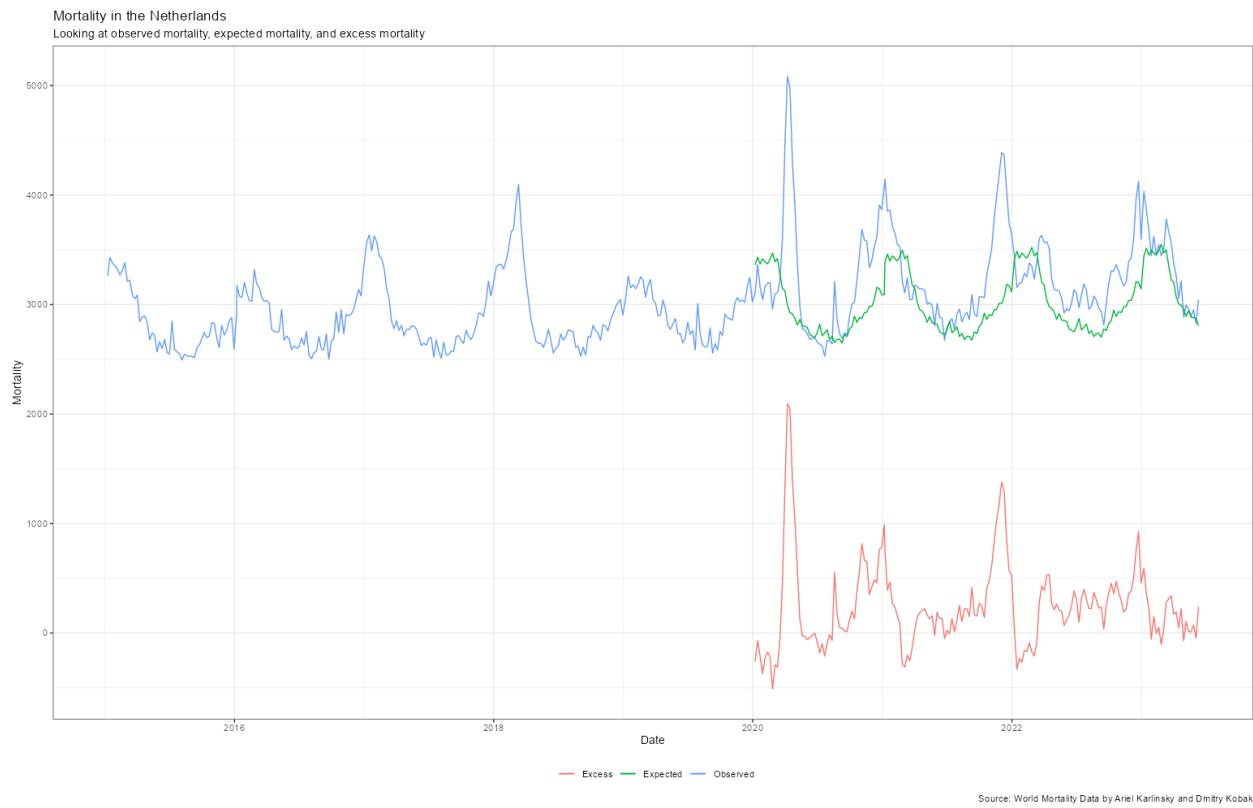


Figure 10. Mortality in The Netherlands, looking at observed, expected and excess mortality. Excess mortality data is not available prior to 2020.

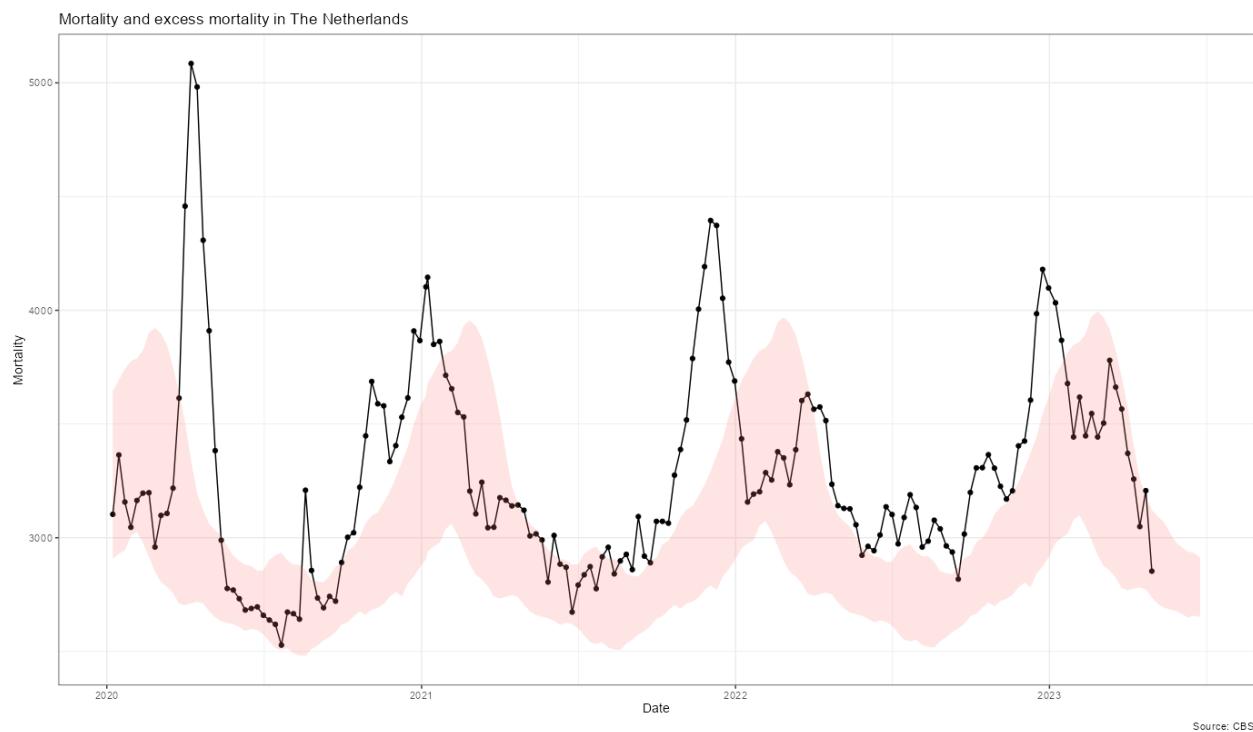


Figure 11. This figure shows the excess mortality in The Netherlands from 2020 onwards. In red is the 95% boundary of the expected mortality which shows a very stable cyclic pattern.

Below is a graph (Figure 12) in which we compared the data coming from the World Mortality Dataset (WMD) Our World in Data (OWID), and CBS. There are slight differences, but in essence they seem to show the same patterns.



Figure 12. Comparison of excess mortality metrics (number and percentage) across three sources (CBS, WMD, and OWID). The results are NOT the same, but they do follow the general line. The differences, although present, seem negligible.

Seeing that the data sources show similar figures, we can now look at the CBS data more closely by splitting expected, observed, and excess mortality by age (Figure 13). Clear to see is that access mortality clearly plays a role from 65 onwards, except during the end of 2022 in which the excess mortality for the 0-65 group was even higher than March 2020.

We can split that data even further into long-term healthcare users (WLZ) and non-WLZ (Figure 14). It seems as if the excess mortality for WLZ is higher than for non-WLZ as a function of a higher expected mortality. Because we do not know how exactly how expected mortality is calculated, we can only infer from that the data that healthcare uses plays some role, but not exactly to what extent and via which pathway. In essence, both groups showed similar mortality patterns.

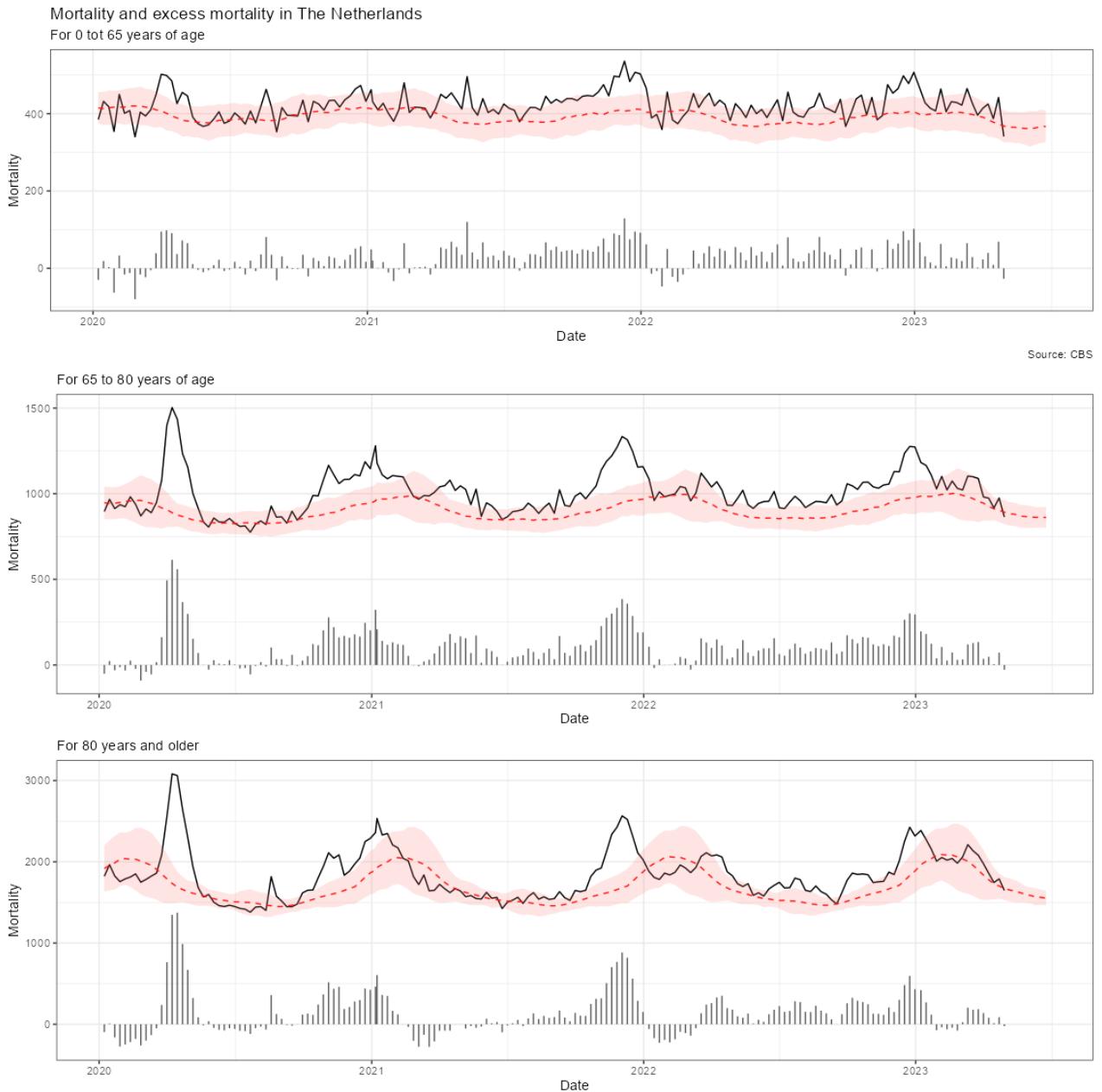


Figure 13. Expected mortality (red line), observed mortality (black line) and excess mortality (bar chart) per age category.

In figure 15 we show the excess mortality as a function of age compared to projection. What is shown is the percentage difference between the reported number of weekly or monthly deaths in 2020-2023 and the projected number of deaths for the same period based on previous years. The black line is the overall excess mortality. Hence, what we wanted to show is how much age plays a role in determining excess mortality. For instance, if the overall excess mortality line strongly mimics the excess mortality line of the oldest population, it is safe to say that most of the excess mortality is coming from the oldest population. Also, the excess mortality lines per age category give us some hints on the

robustness of the model. Heavy fluctuations in excess mortality hint at an unstable and perhaps even useless model.

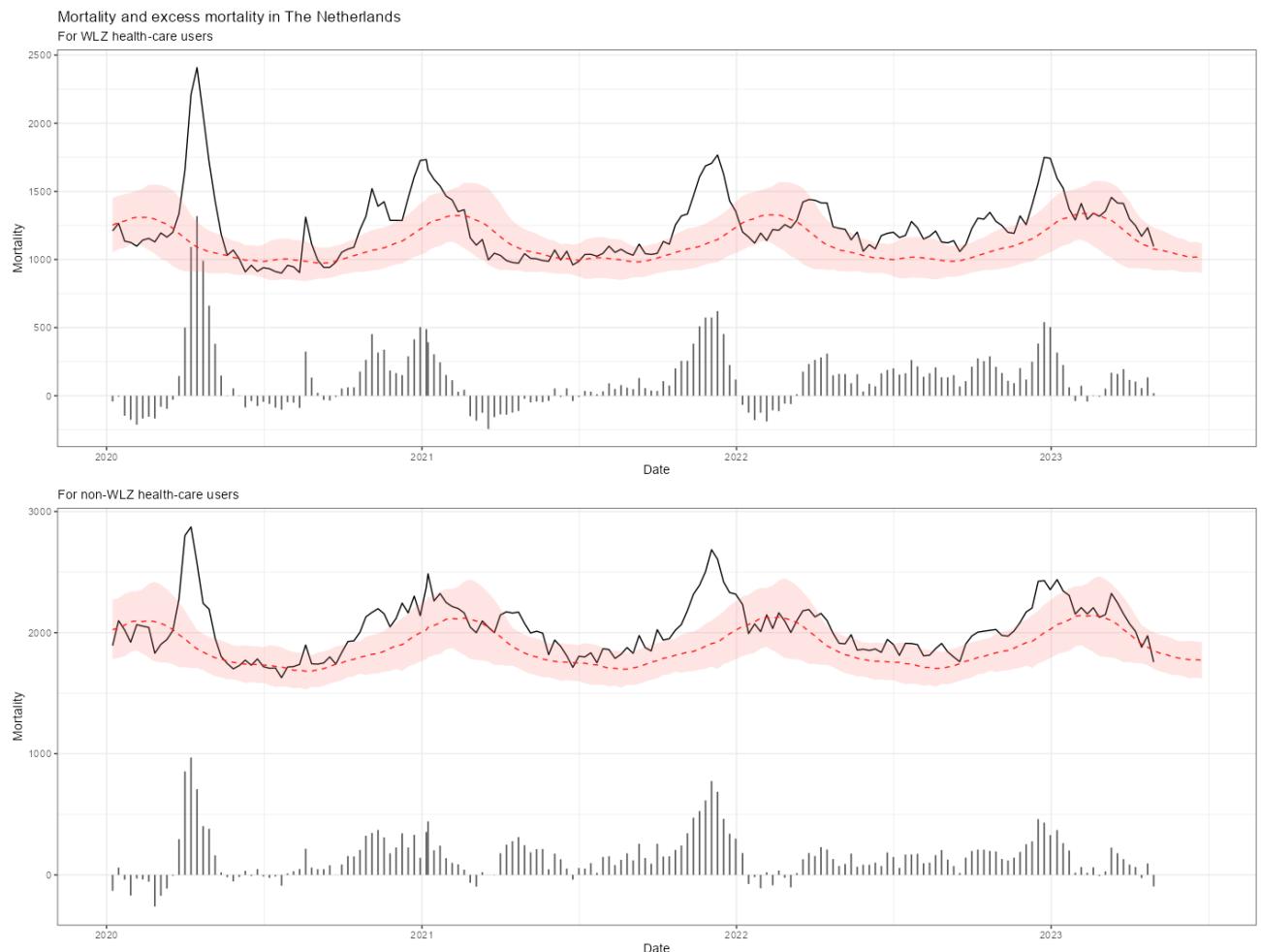


Figure 14. Mortality and excess mortality (in red) for health care and non-health care users. The same pattern is clear to see but differs somewhat in the numbers of death. Hence, why for health-care users there are three clear spikes that are extent beyond the upper limit of the projected mortality.

Clear to see in Figure 15 is how the 0-14 age category fluctuates heavily, showing the limits of the metric as a function of age. The rest of the average excess mortality is made up of categories 15-65, 65 to 74, and 75-84. The first category is the broadest which limits its usefulness. Furthermore, this picture clearly shows the limitation of the metric of excess mortality. Because it is a normative value, you can either be below it or above it, but you are unlikely to be on it. In the absence of a measure of clinical relevance, a number on the excess mortality metric (either below or above) has limited relevance.

Sometimes, the excess mortality of an age category is lower or higher than the average, but this is to be expected considering the way the metric is calculated (expected vs observed). Hence, it is very difficult to use the excess mortality metric to infer any form of pathway to

the role of age. Except for the youngest and the oldest category – the youngest does not match the metric, whilst the oldest category had many spikes like the average.

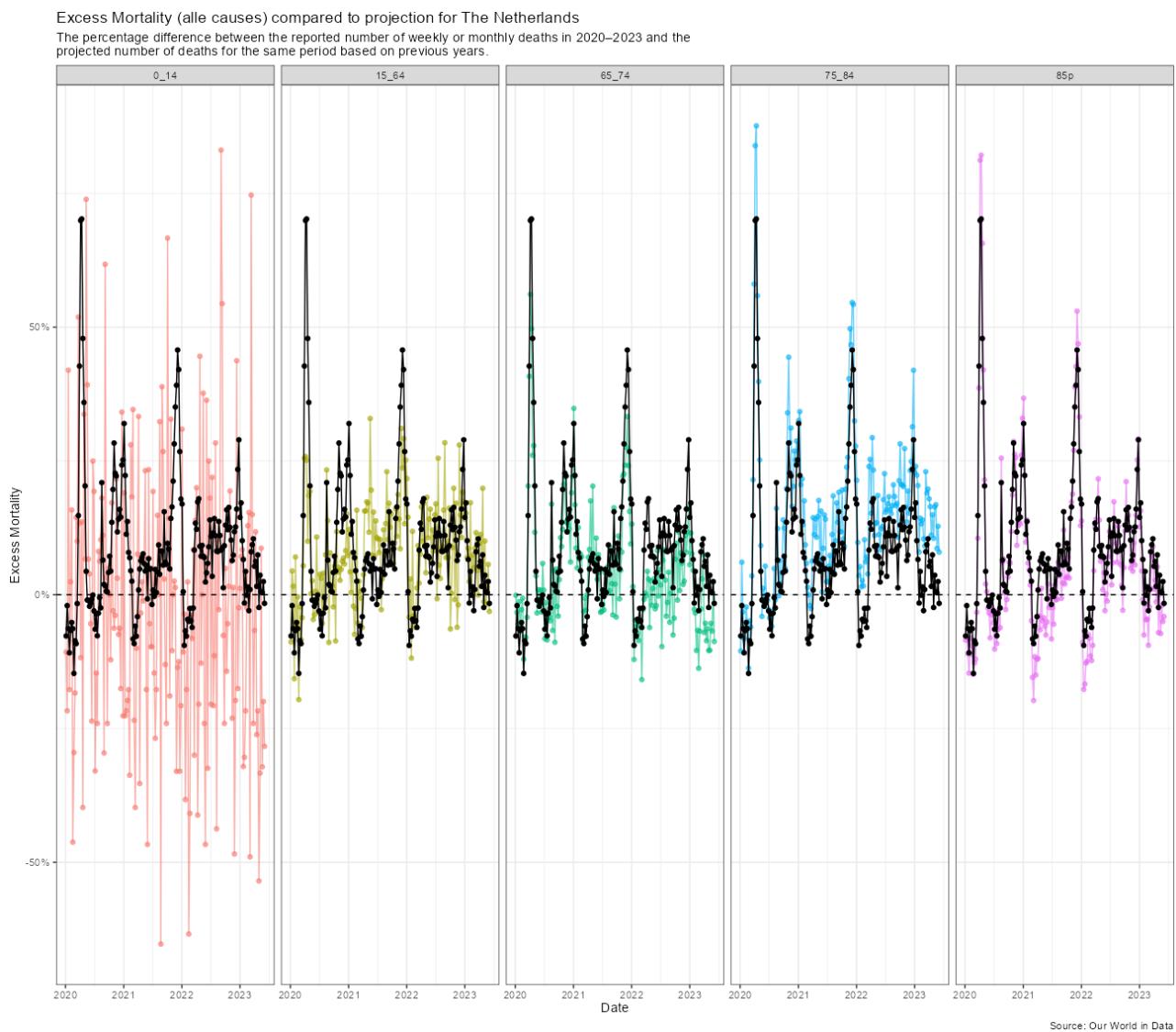


Figure 15. This figure shows the same data as the figure before but split per age category. This way, you can clearly see the average excess mortality across all data (black line) and the excess mortality per age category.

We have discussed several times how much excess mortality is determined by the expected mortality, but we do not know how the expected number is being calculated. In the OWID database, there is a dataset that shows two different models which lead to two different excess mortality calculations. Figure 16 is perhaps one of the most important plots of this entire document, showing the excess mortality for the Netherlands using WHO and Economist data which is made available via OWID. Depending on the model used to calculate projected mortality, you will also find a difference in excess mortality. Considering that one of the goals of this project is to estimate if excess mortality has a cause in late-onset cancer screening and care, it seems that it does make sense which source you use. For

instance, consider January 2021. The WHO estimates around fifteen thousand extra deaths in The Netherlands, whilst the Economist estimates around seventeen thousand. These 2000 only exists in the world of models but need to be accounted for in the real world by looking at the source of death (which is only available for real deaths). Hence, since we are estimating the possible contribution of cancer care during the Covid-19 pandemic on excess mortality, the number provided is absolute key. But no absolute number can be provided.

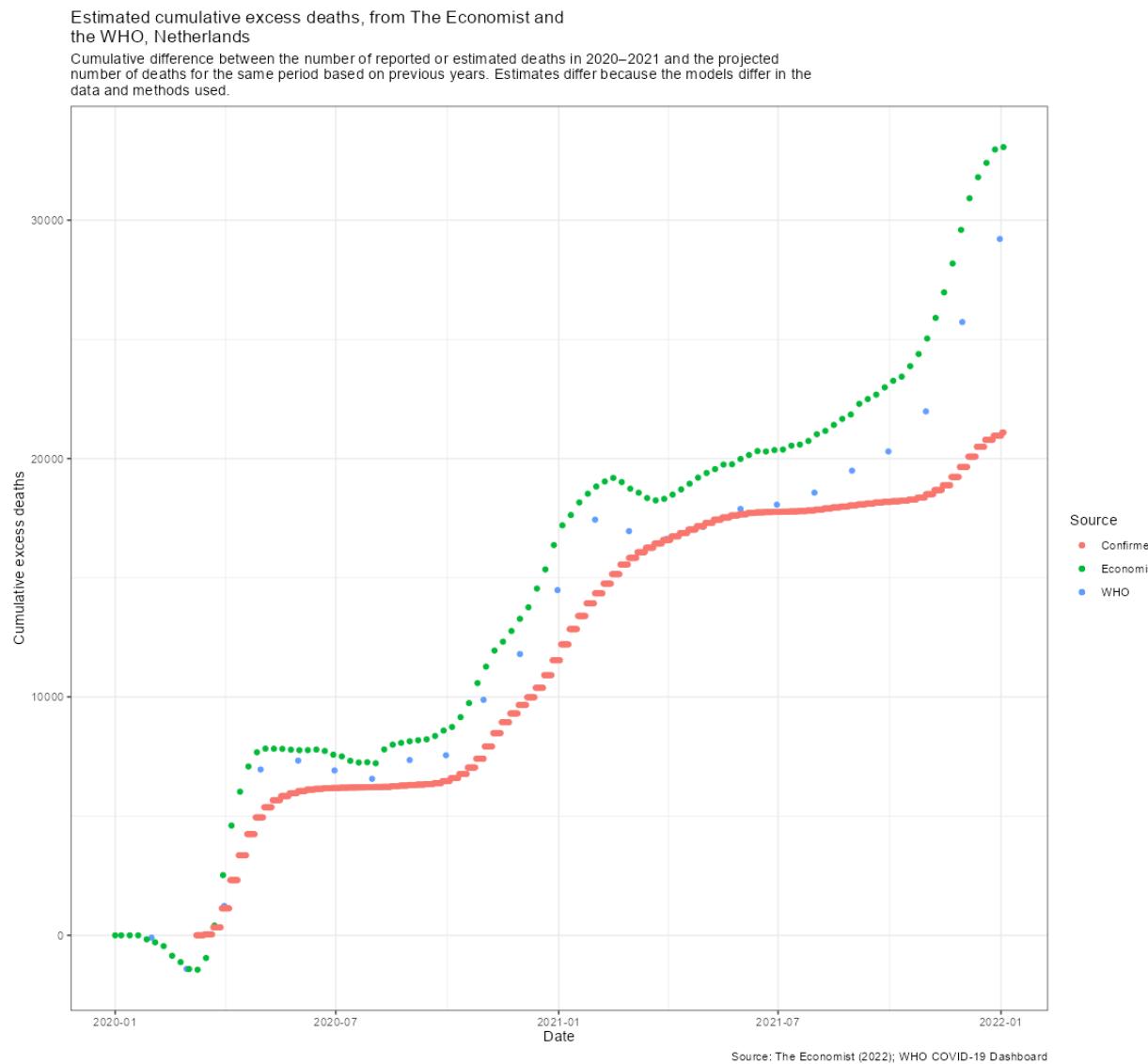


Figure 16. This graph shows the excess mortality for the Netherlands using WHO and Economist data which is made available via OWID. Depending on the model used to calculate projected mortality, you will also find a difference in excess mortality.

Cause of death

From [life expectancy](#) to [mortality](#) onwards to [excess mortality](#) it is now time to discuss the causes of death. Of course, the entire aim of this paper is to look for the possible pathway of delayed care in cancer patients towards excess mortality. The largest provider of data on causes of death is the CBS. In Figure 17 we show the total number of deaths per age category across the years in the Netherlands. As a heatmap, one can clearly see that the last three years were amongst the deadliest in the past 75 years, especially for men.

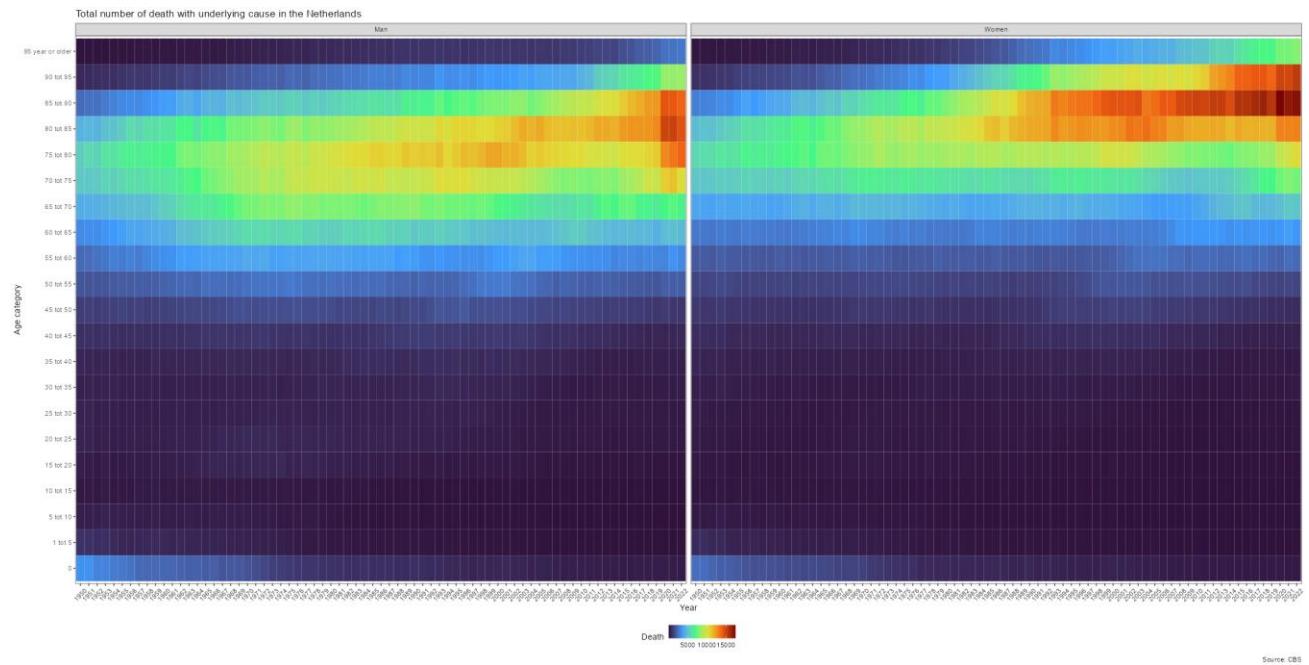


Figure 17. Total number of deaths per age category across the years in the Netherlands.

The total mortality has been rising for years, especially for women, which is most likely a direct consequence of our ageing population in combination with a higher life expectancy in women.

Figure 18 shows that the number of deaths for which not a specific cause of death is described has also increased, especially in the past three years. The highest number of unclearly described causes of death can be found in males, aged 75 to 80, in 2022.

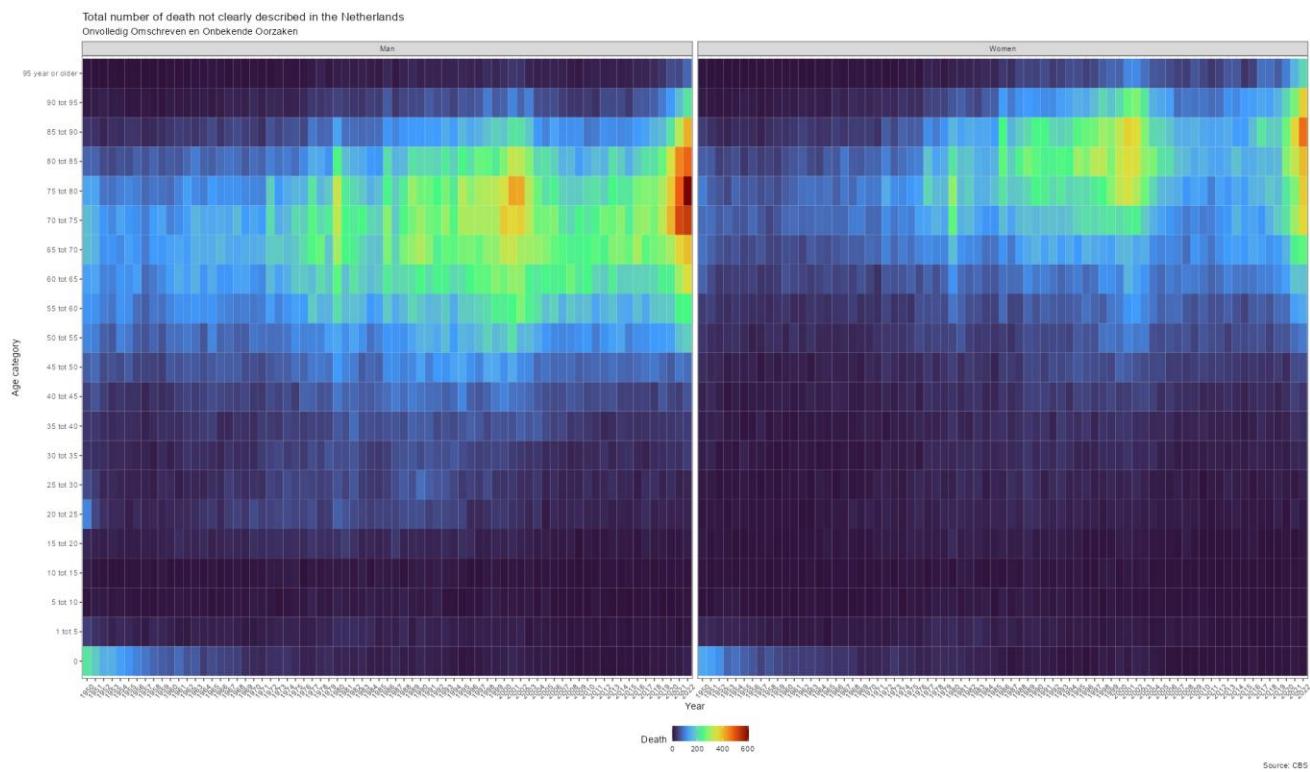


Figure 18. Total number of deaths not clearly described in the Netherlands.

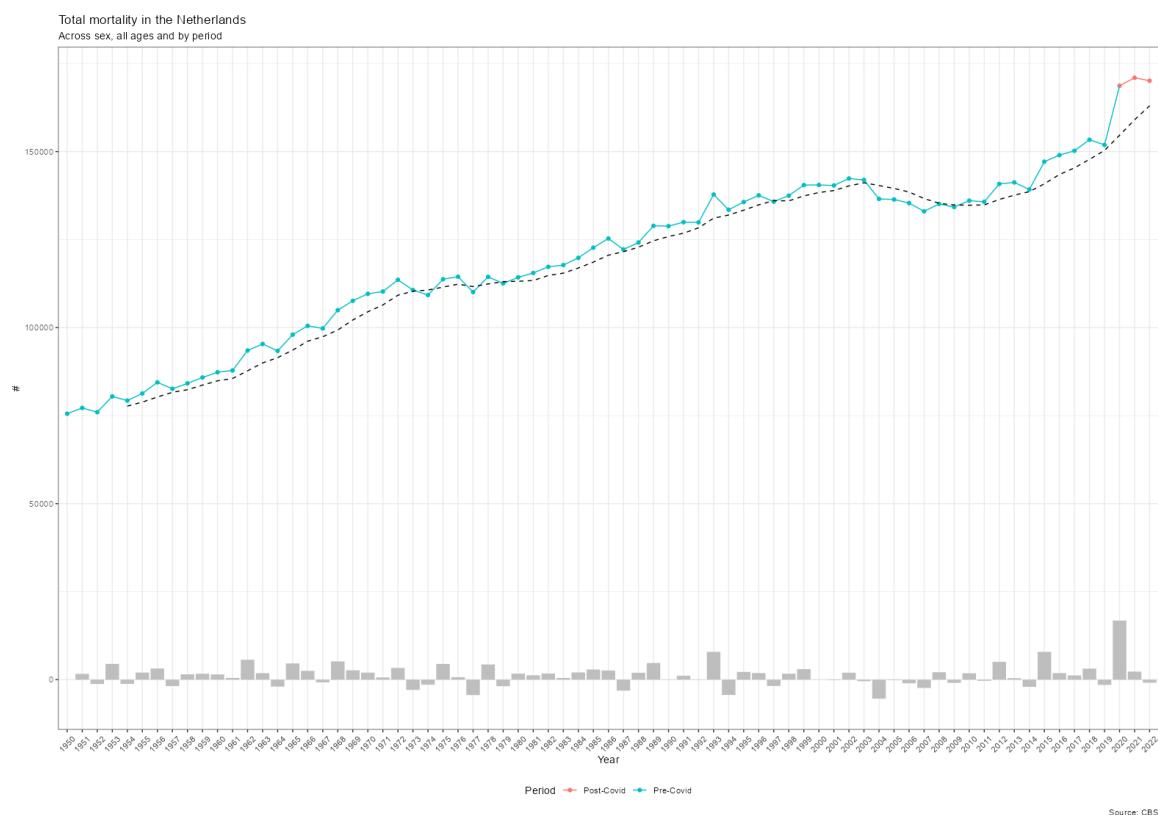


Figure 19. Total mortality in the Netherlands across sex, all aged and by period (pre- vs post-Covid). The black dotted line is a smoothed line coming from a LOESS estimator and is only added to showcase a trend based on the observed data. The bar charts are differences in mortality from one year to another year. The highest bar chart is 2020 which shows the steepest incline in total mortality.

Figure 19 shows that the past three years have shown the highest mortality numbers with the strongest increase in 2020. The year 2021 showed an even higher total mortality after which there was a decline in 2022. Nevertheless, all three years show a considerably higher mortality rate than the previous years. This cannot be explained by increased ageing, life-expectancy or population growth.

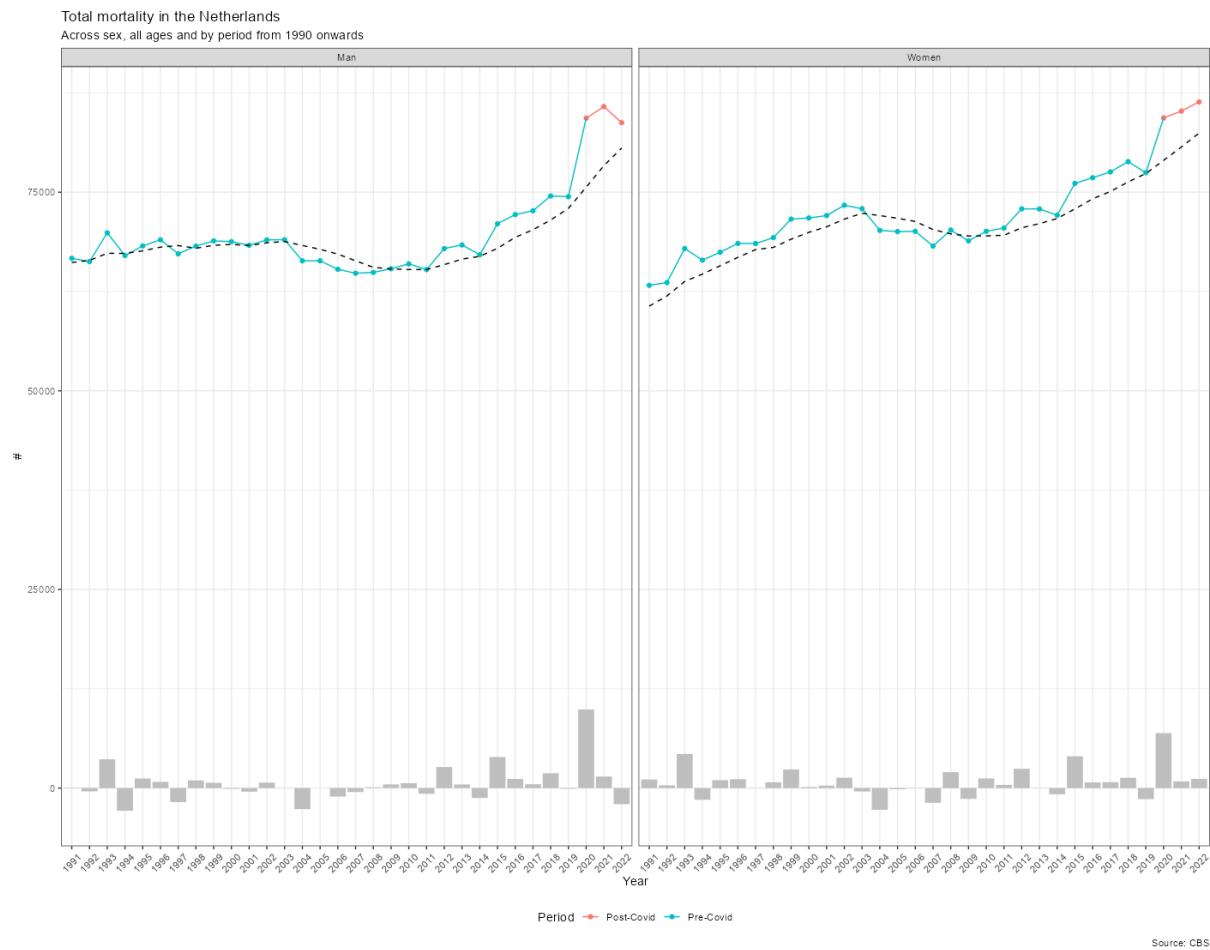


Figure 20. Same graph as figure 19 but then show per sex.

Figures 20 shows the same data as Figure 19 but then split by sex. Clear to see is the overall trend of mortality increase in 2020, but the development differs from there. For instance, where male death decreased in 2022, female death increased making 2022 the deadliest year based on raw numbers.

Of course, Figure 19 and Figure 20 do not show the underlying cause of death, but what they do show is that mortality changed considerably in the past three years. In the next graph (Figure 21) we show the underlying causes of death for all major categories as provided by the CBS.

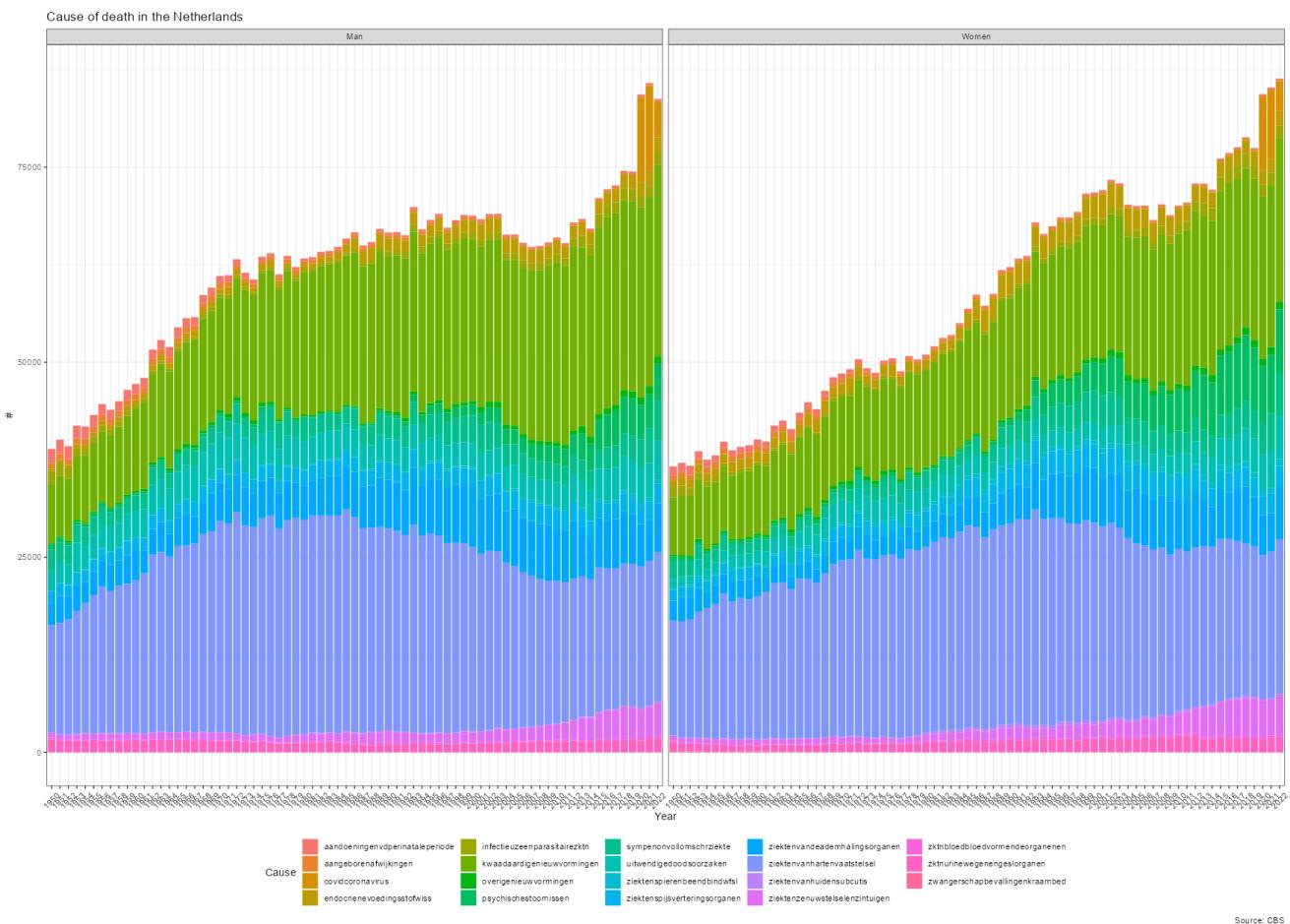


Figure 21. Causes of death per sex, and year, in the Netherlands.

Figure 21 shows per major category how much they contributed to the total mortality numbers. There is a special category (Covid-19) which only exists from 2020 onwards. Because our focus is on cancer mortality during the Covid-19 period, we will only focus on these causes of death. To compare the impact of Covid-19, we will also look at the flu since this is the infectious disease that has traditionally led to delayed care in the Netherlands¹⁰⁴. The year 2018 was especially bad¹⁰⁵ and so we will focus on this year as well. From the (excess) mortality data, we could already see a spike in 2018.

Figure 22 shows the total mortality split by cause of death being Covid-19 or not. Clear to see is how Covid-19 was attributed a relatively large part of the causes of death for 2020 and 2021, after which the number dropped. We see somewhat similar numbers for males and females.

¹⁰⁴ https://www.limburger.nl/cnt/dmf20180228_00056816

¹⁰⁵ <https://www.bd.nl/binnenland/spoedeisende-hulp-vol-met-griepatienten~a822b56a/?referrer=https%3A%2F%2Fwww.google.com%2F>

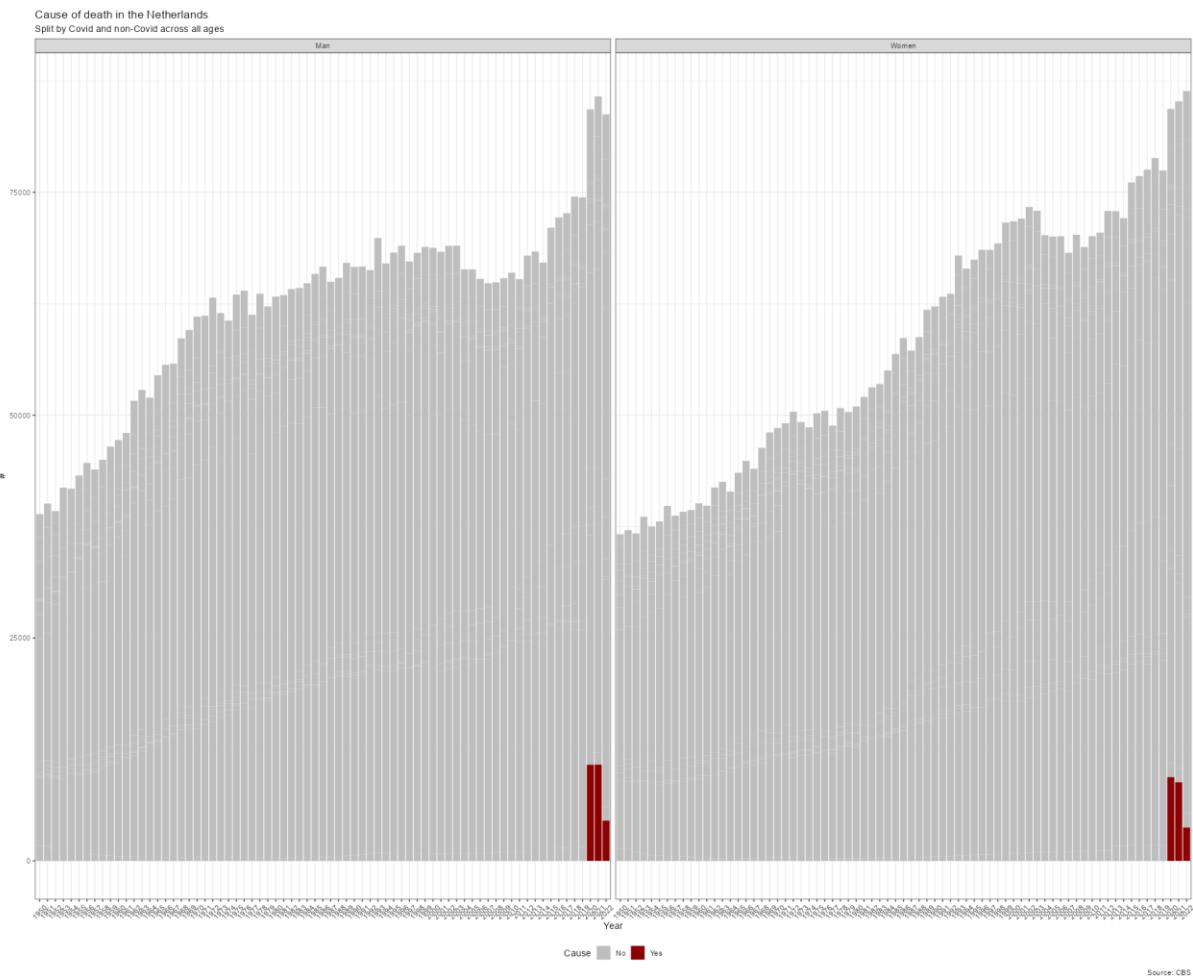


Figure 22. Cause of death specified by Covid or not.

Comparing Covid-19 to the flu clearly (Figure 23) shows that for the 50 years or so, the flu has contributed little to the cause of death. The excess mortality seen in 2018, which is second only to the 2020 spike, is unlikely to be caused by the flu if we look at the numbers on the causes of death.

Figure 24 shows that compared to cancer both the flu and Covid-19 are small contributors to the overall causes of death.

In Figure 25 we show the same data as in Figure 24 but we leave out the rest. Once we do that, we can see that cancer as cause of death has been on the rise for years at fluctuating speed. In men, cancer has decreased as a cause of death from 2018 onwards with a small increase in 2022. The flu has really been a minor contributor over the years, and the part attributed by Covid-19 is also decreasing again.

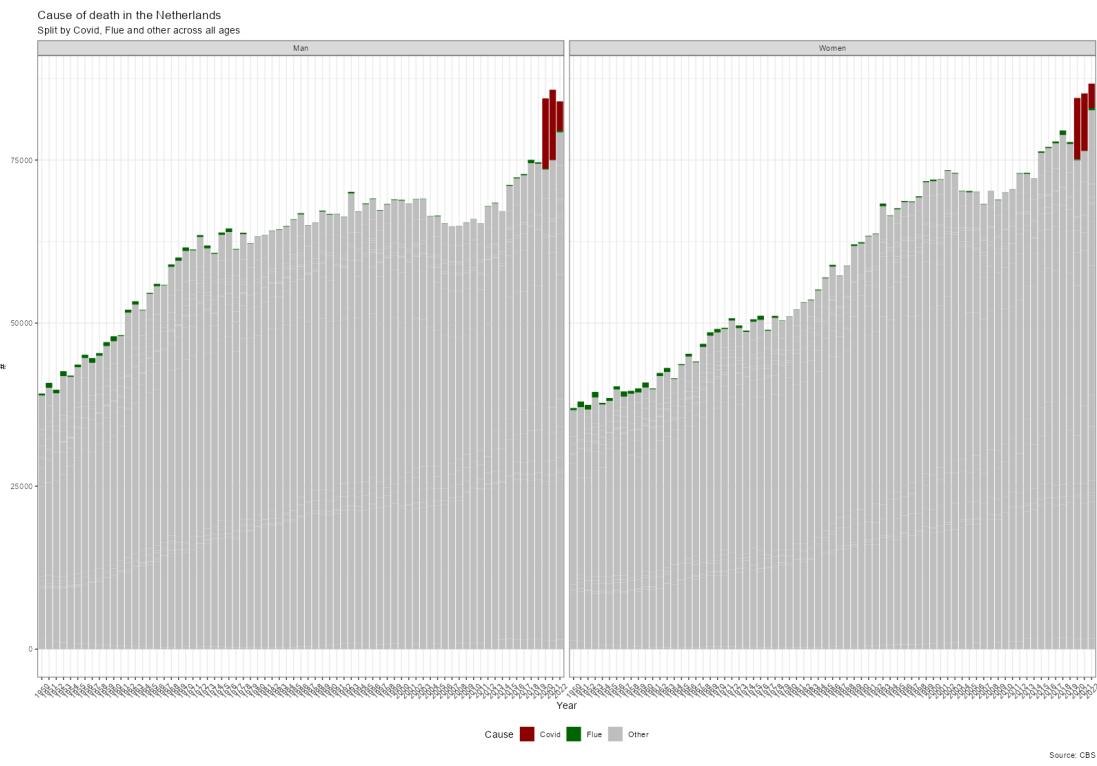


Figure 23. Cause of death being Covid-19 or the flu.

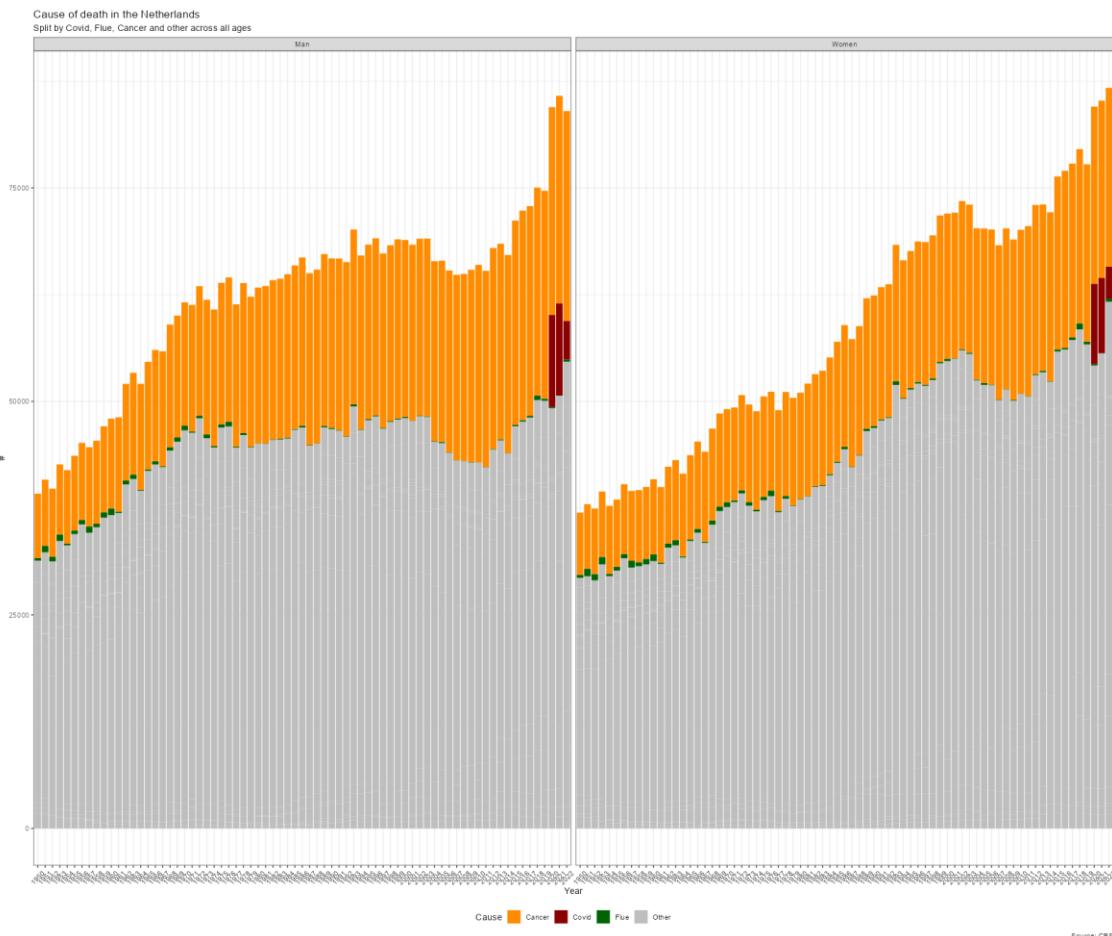


Figure 24. Causes of death with a focus on cancer, Covid-19, and flu.

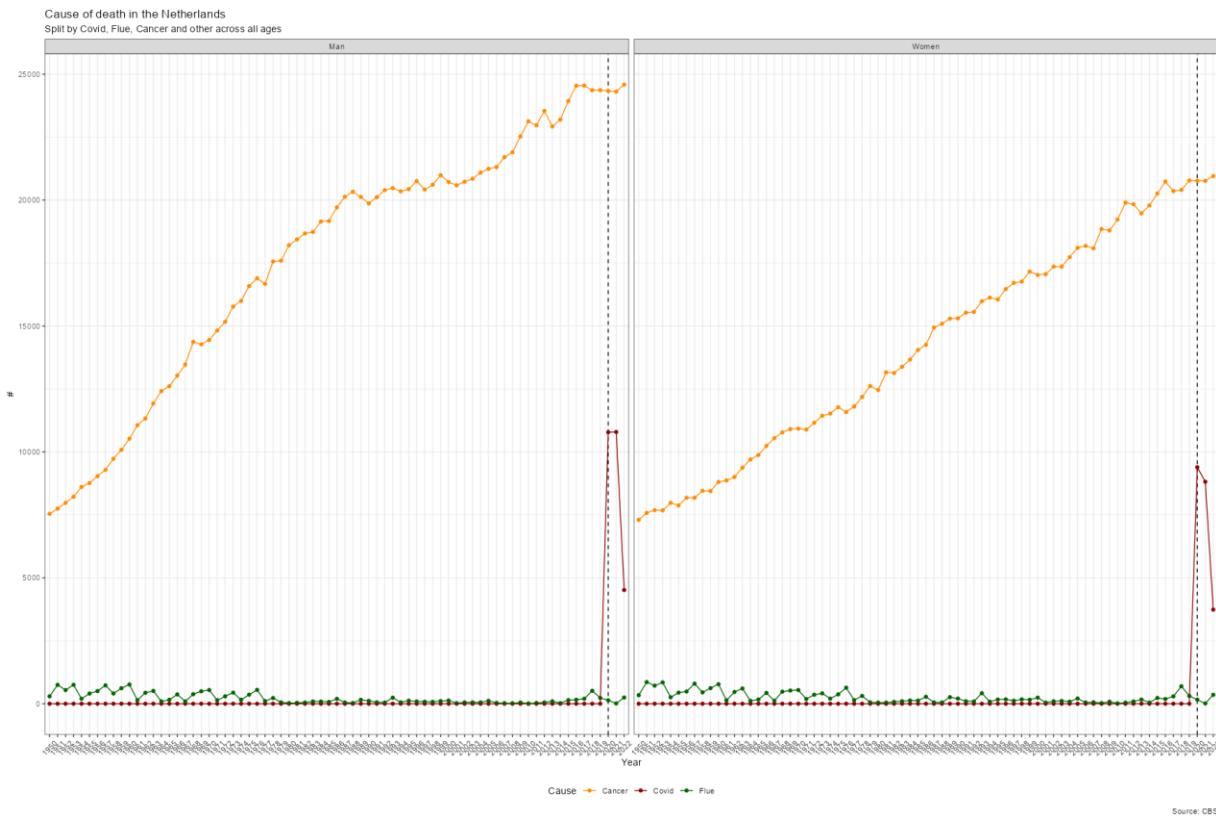


Figure 25. Causes of death by cancer, Covid-19, and the flu.

Covid-19

It is now time to focus more specifically on Covid-19. Although the causes of death sources coming from CBS already show the attribution of Covid-19, it is good to dive a bit deeper. When assessing the possibility of excess mortality via delayed care, we must also delve deeper into the most plausible contributor to excess mortality which is Covid-19 itself.

There is a difference it seems between confirmed and probable Covid-19 death which seem to play a role in 2020 and early 2021 (Figure 26). For the majority, all registered Covid-19 deaths are confirmed Covid-19 deaths.

According to the data of CBS, Covid 19 was a heavy contributor to the excess mortality as can be shown in Figure 27.

If we go deeper into the Covid-10 data, splitting it by healthcare use (WLZ), age and gender, we see that there is no real difference in patterns based on healthcare use (Figure 28) and many people who died of Covid-19 are older than 65 (Figure 29). In the 65 - 79 range, more males died but it is difficult to disentangle this from ‘normal’ life expectancy (Figure 30).

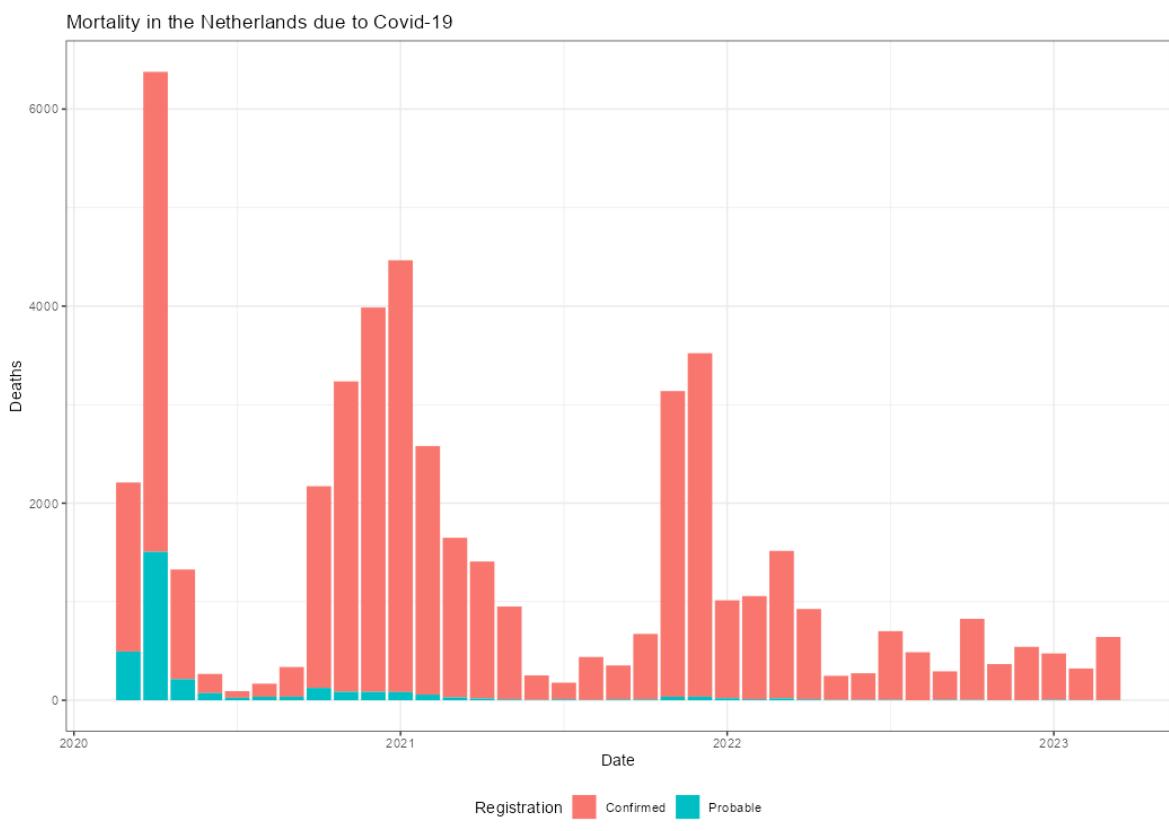


Figure 26. Mortality in the Netherlands due to Covid-19.

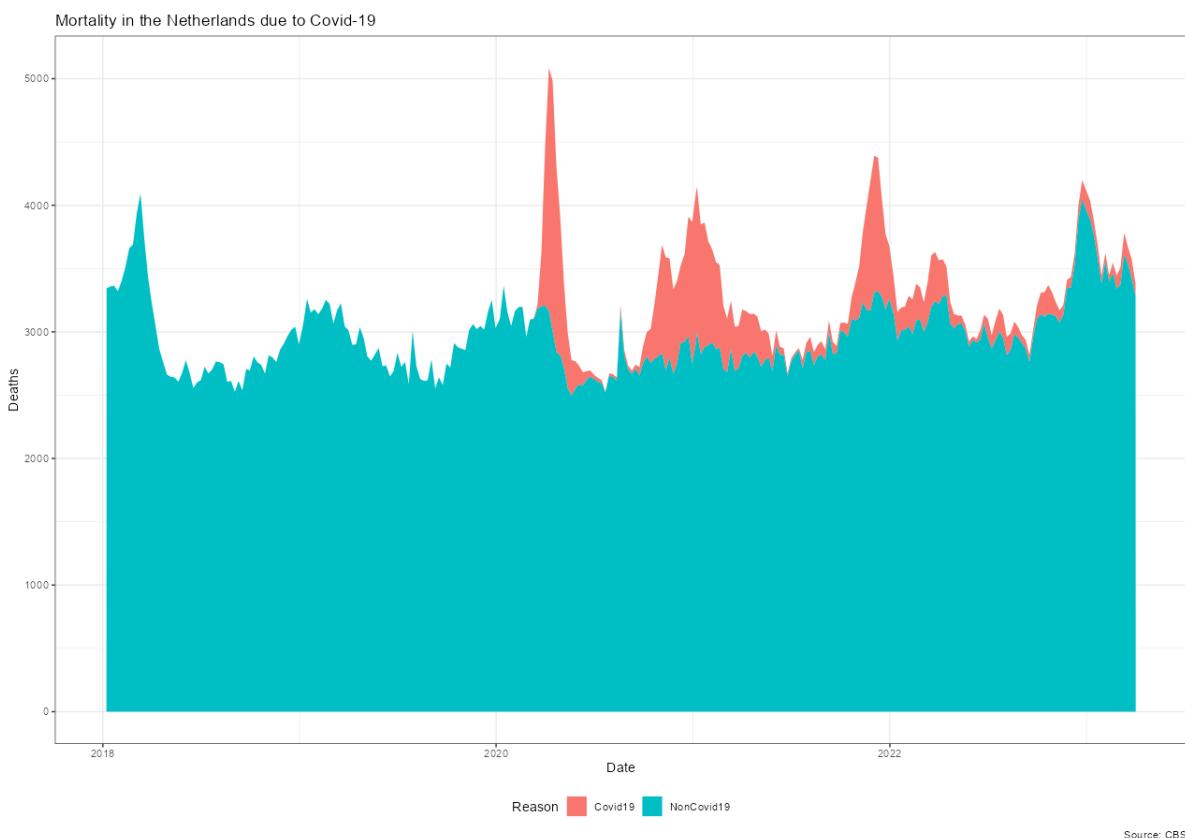


Figure 27. Mortality in the Netherlands attributed to Covid or non-Covid.

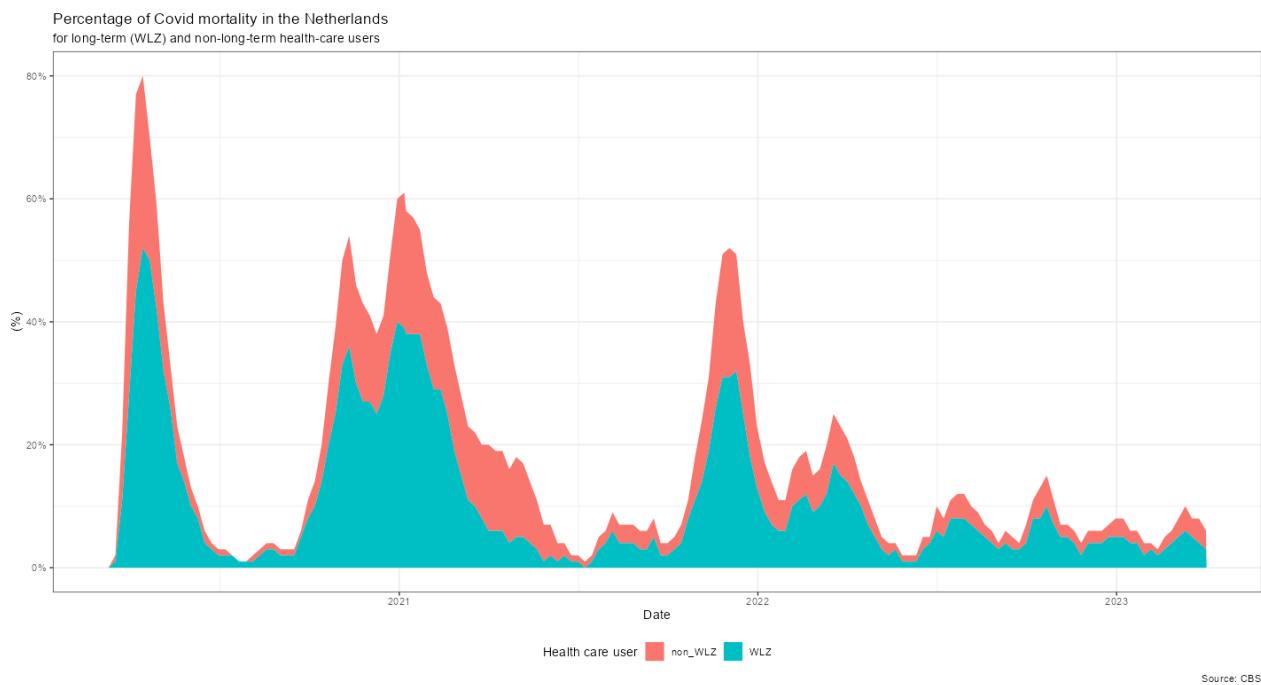


Figure 28. Percentage of Covid-19 mortality based on long-term vs non-long-term healthcare users.

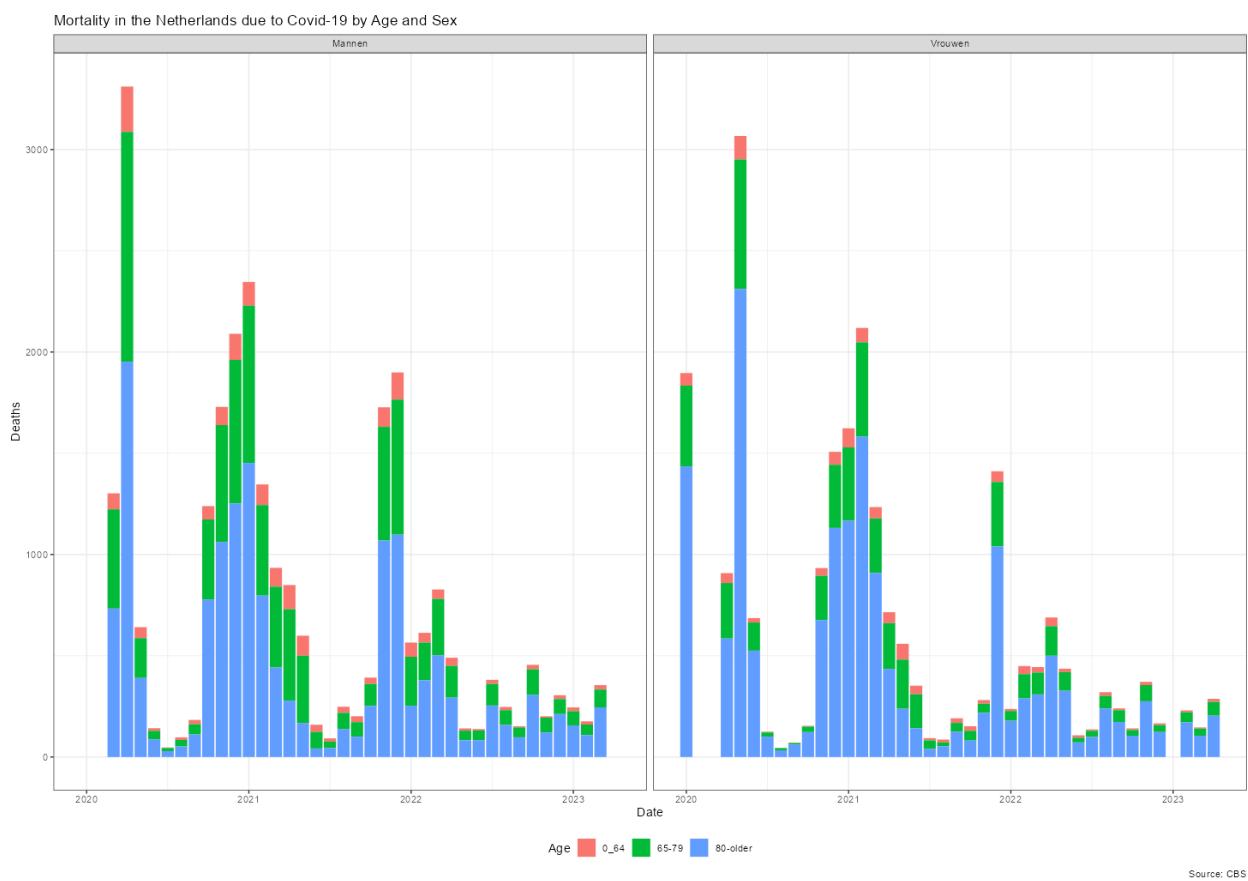


Figure 29. Mortality in the Netherlands due to Covid-19 split by age.

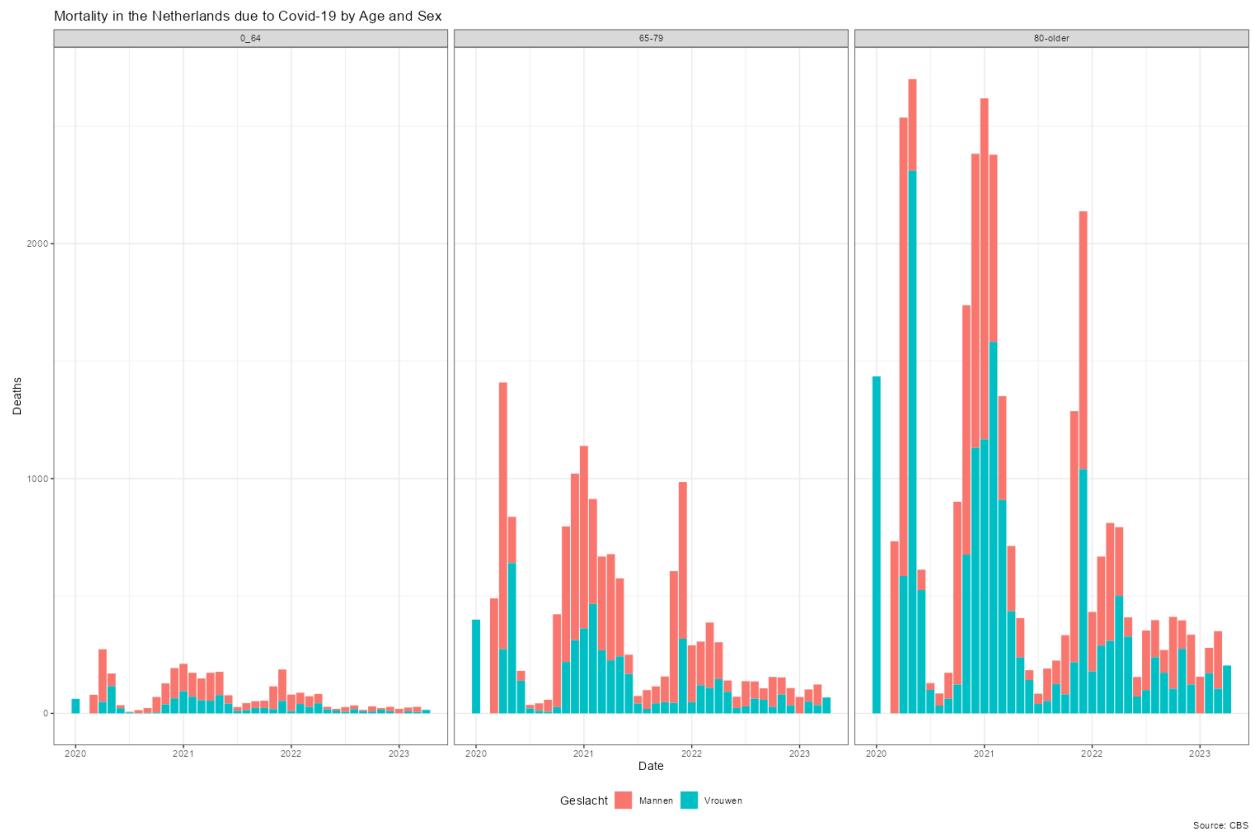


Figure 30. Covid-19 mortality split by age and sex.

In summary, it seems that Covid-19 mortality does not show strange patterns. The older somebody is, the higher the likelihood of dying from Covid-19.

Cancer

We already showed that cancer mortality was still much higher than Covid-19. For years, cancer has contributed tremendously to the overall mortality next to cardiovascular disease.¹⁰⁶ So, it is wise to look deeper into cancer in general as a cause of death and the three types of cancer that interest us the most: breast, cervical and colon cancer.

Figure 31 shows the raw number of cancer deaths from 1950 onwards per age category and sex. It is clear to see that more men than women die of cancer and one can also see that cancer mortality is rising. We already showed this in the previous images. For men, 2022 seems to have the highest rate of cancer especially in the 75 – 80 category.

¹⁰⁶ <https://www.cbs.nl/nl-nl/longread/statistische-trends/2021/doodsoorzaken-2000-2020?onepage=true>

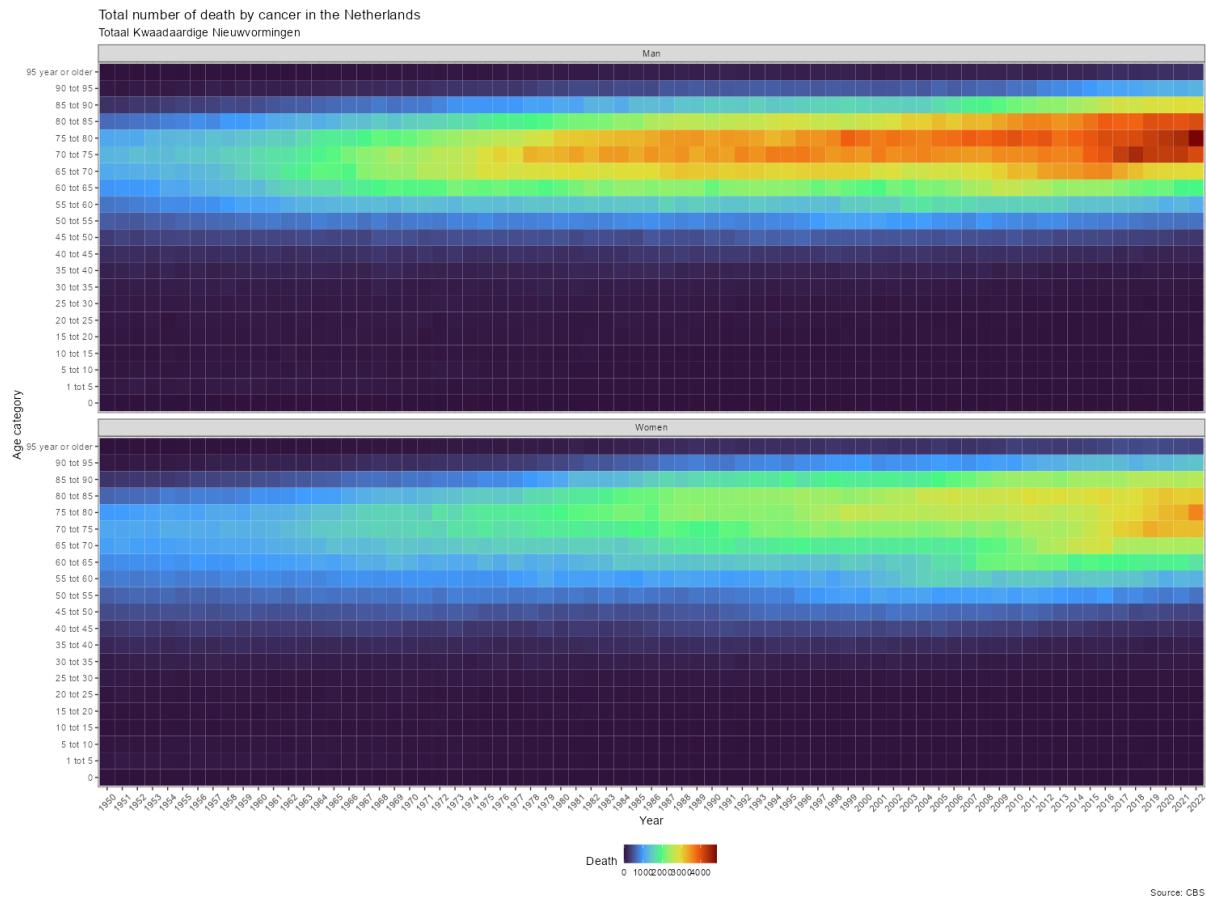


Figure 31. Total number of cancer death by year, age, and sex.

Breast cancer

Figure 32 shows the breast cancer mortality data across years and age. It seems as if a couple of years in the early 2000s showed a decrease in cancer mortality, but overall, the mortality seems to be picking up slowly (Figure 33). Perhaps a better way to show the data is by using standardized numbers which are corrected for the population growth. Coming from a different source ([VZinfo](#)) it seems that breast cancer mortality has been slowly decreasing for years (Figure 34). Unclear is why certain fields did not produce any standardized data. If we go back to the raw data again, using the [NKR-date from IKNL](#), and compare total breast cancer mortality across the years, we see no real difference for 2020 and 2021 compared to the other years (Figure 35). Although some fluctuation is shown, the general trend shows a decrease in breast cancer mortality. If we specify even further by age categories, we cannot see a trend reversal showing severely more or less mortality during Covid (Figure 36).

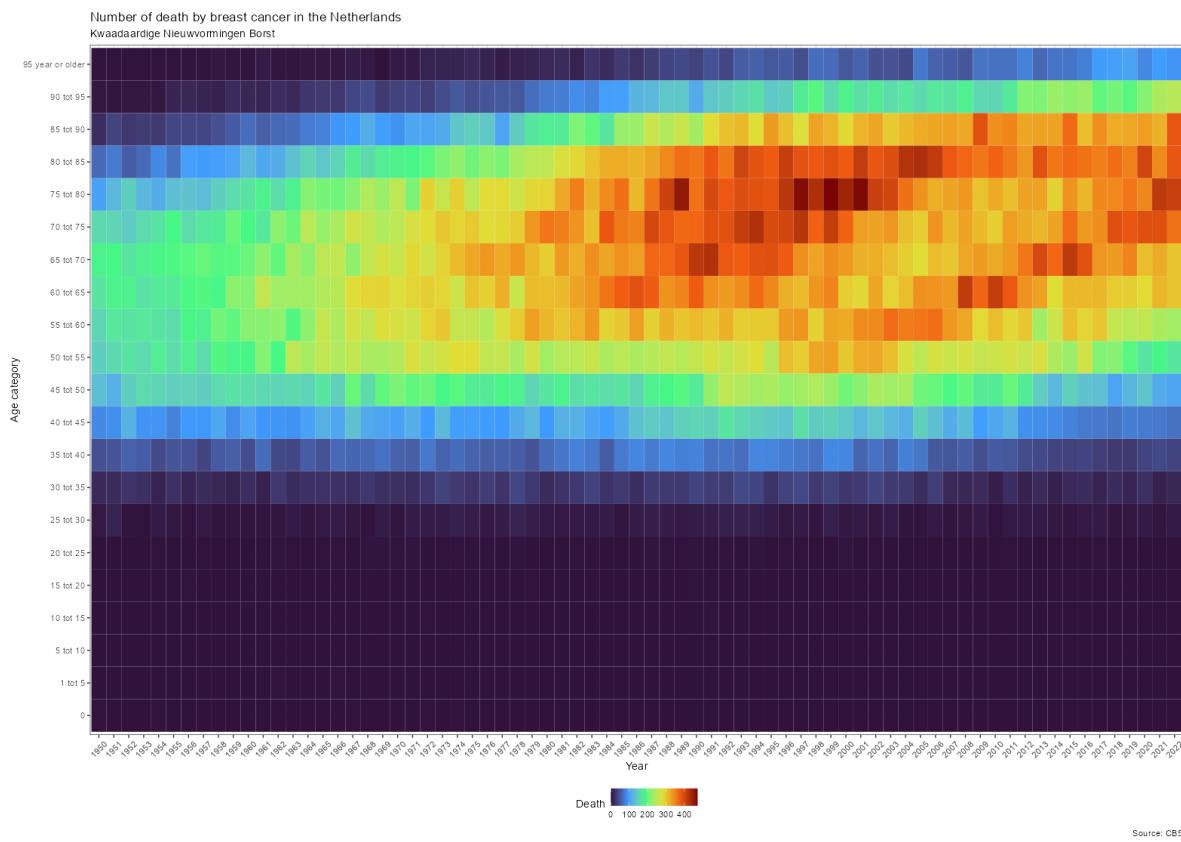


Figure 32. Breast cancer mortality data across years and age.

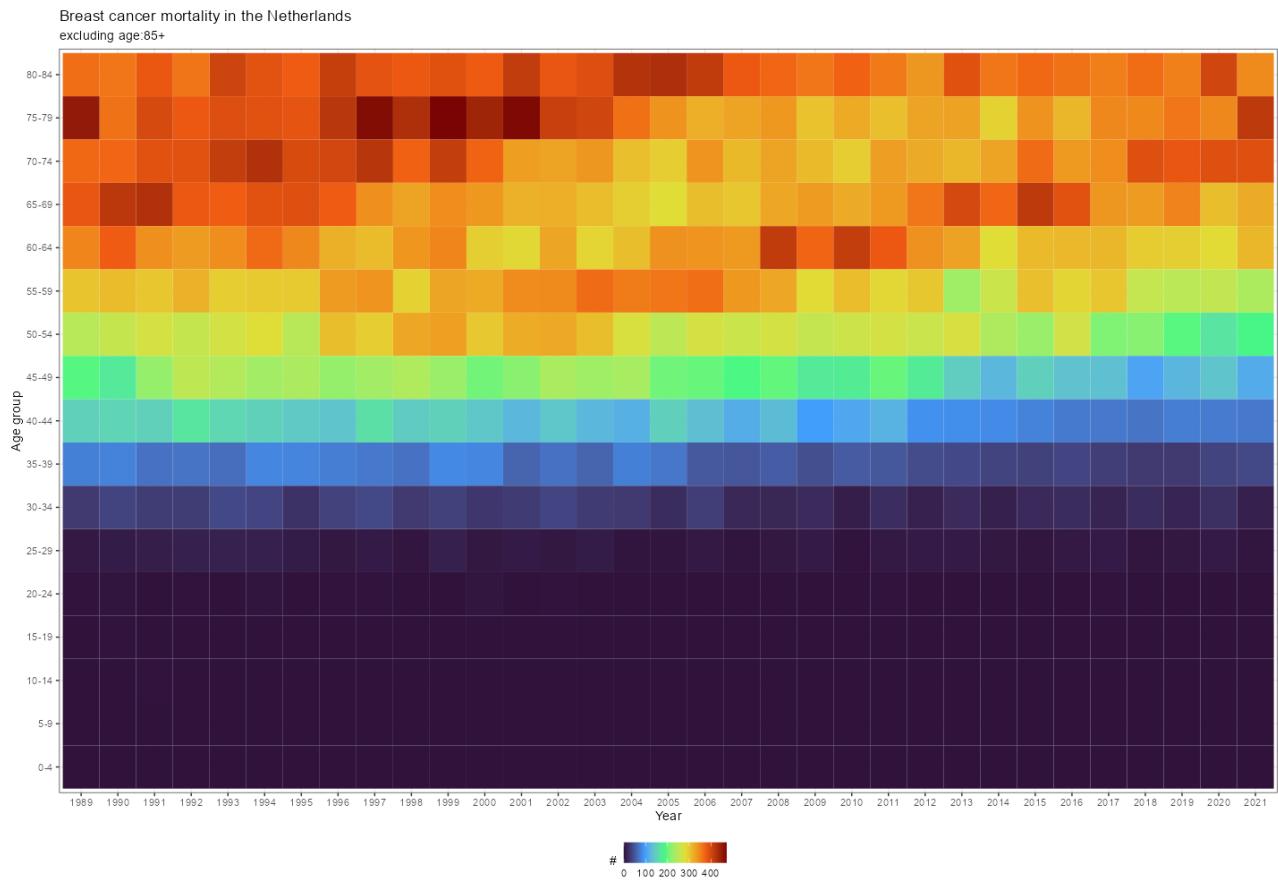


Figure 33. Breast cancer mortality across age groups and years.

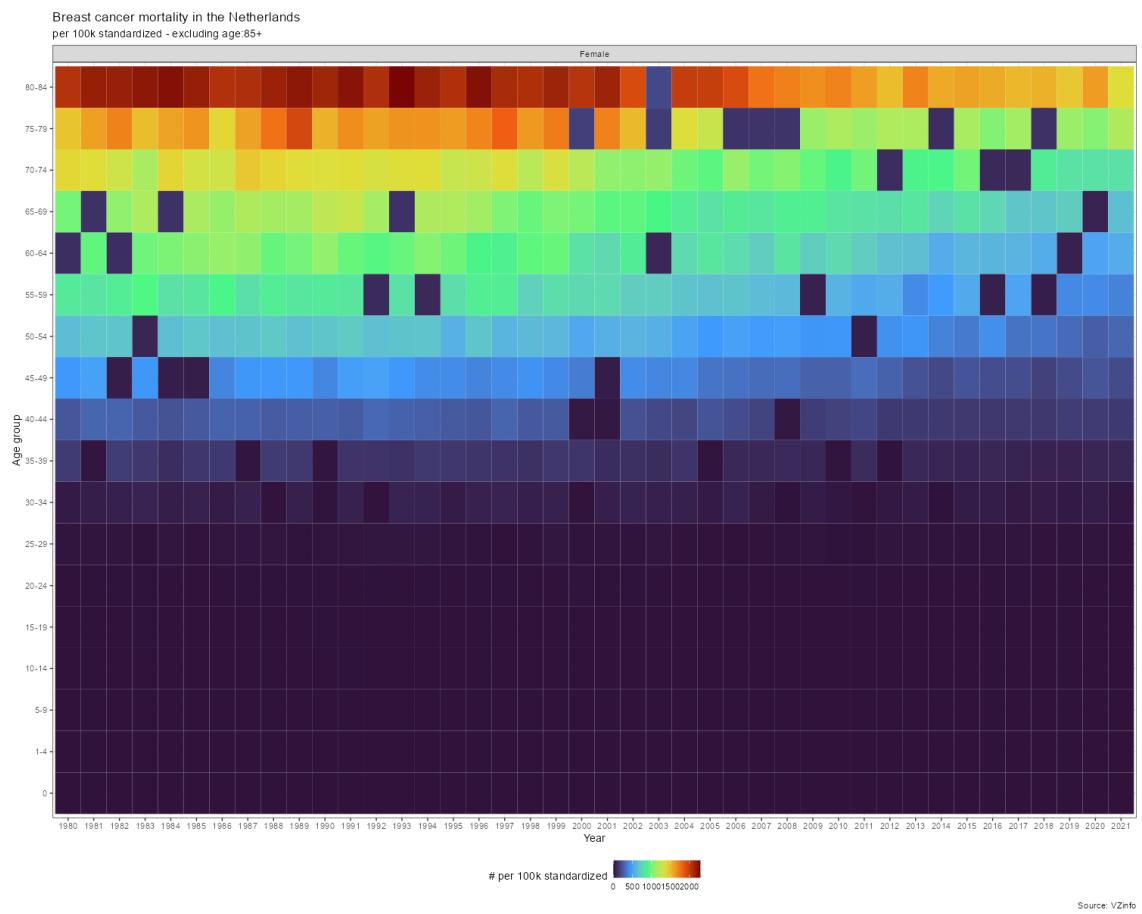


Figure 34. Standardized breast cancer mortality per year and age category.

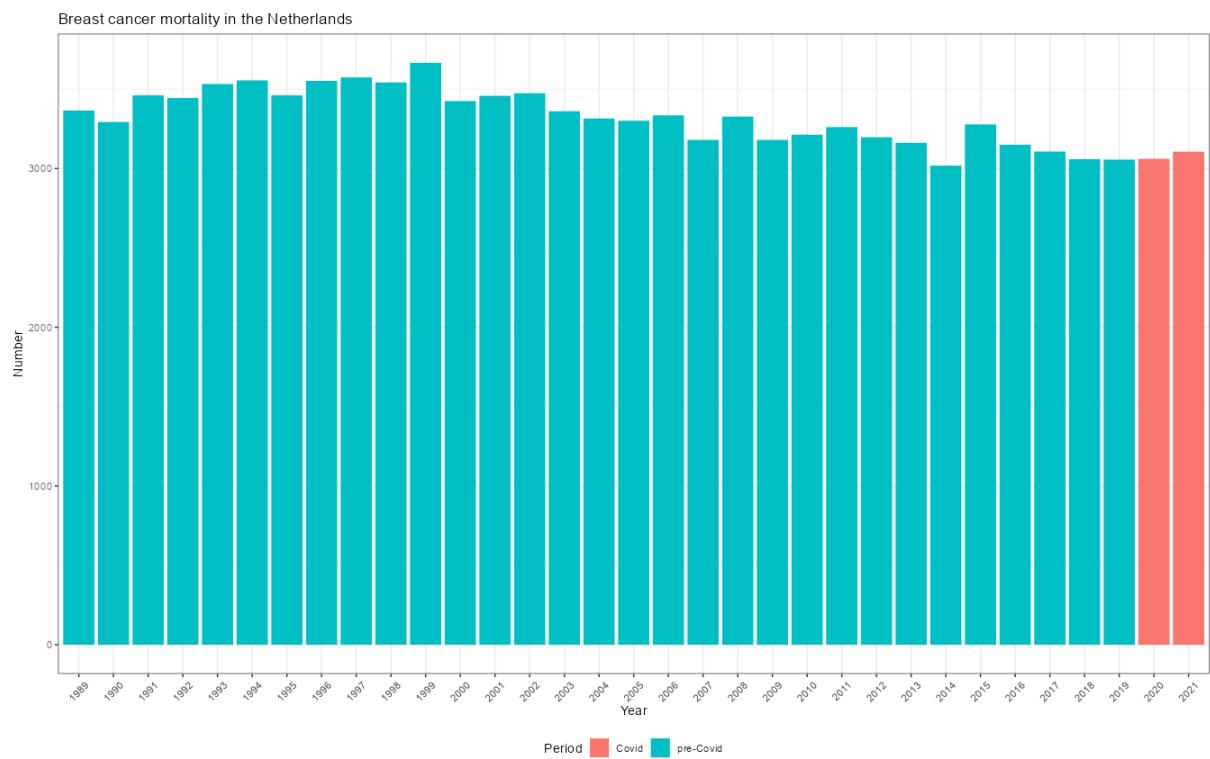


Figure 35. Total breast cancer mortality across the years.



Figure 36. Total breast cancer mortality across years and age category.

Cervical cancer

Cervical cancer has seen quite a decrease in mortality across the years but can already be found in ages as young as below 20 (Figure 37).

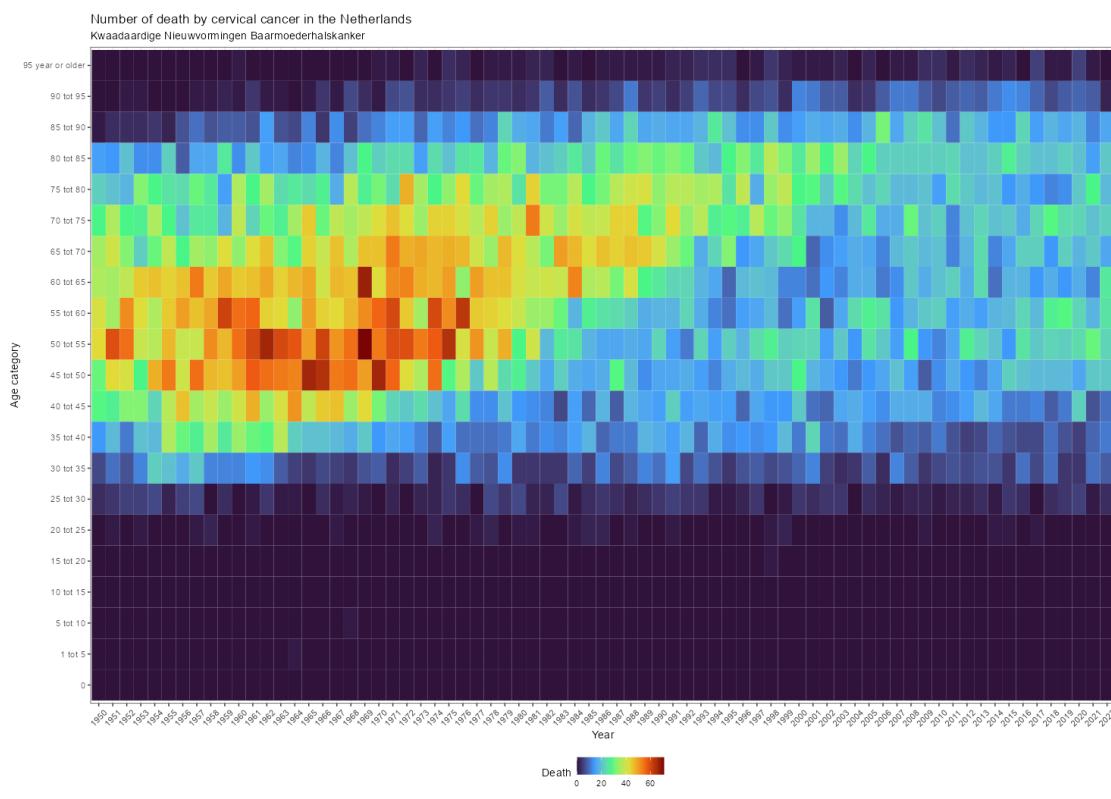


Figure 37. Cervical cancer mortality in the Netherlands from 1950 onwards.

As with breast cancer, the graphs differ somewhat depending on the source you are using. If we look at [CBS data](#) (Figure 37), [IKNL data](#) (Figure 38) and [VZinfo data](#) (Figure 39) it differs somewhat. Especially in Figure 39, which shows three ways of depicting cervical cancer mortality, it shows how unstandardized versus standardized data makes quite the difference.

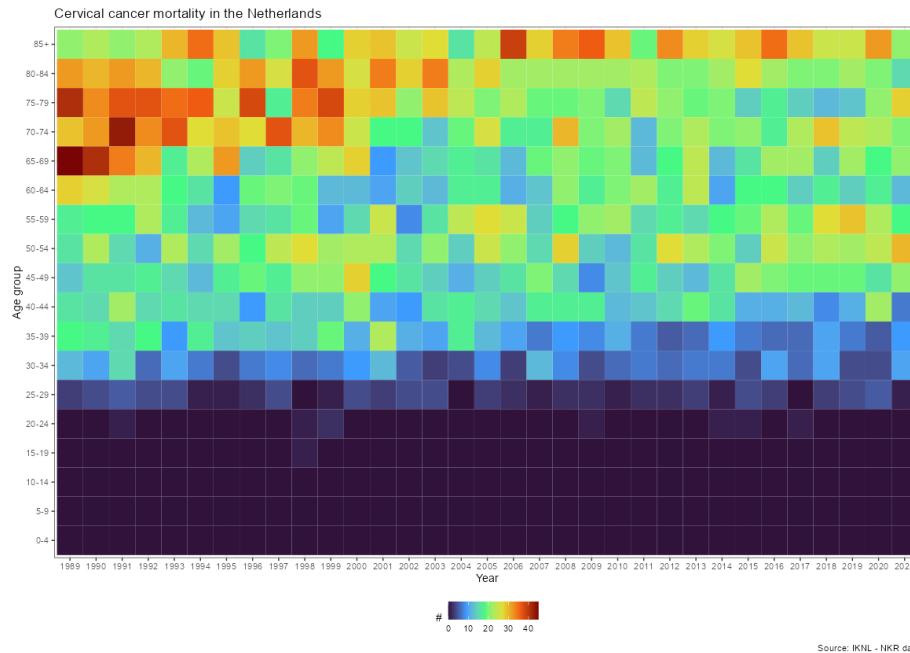


Figure 38. Cervical cancer mortality in the Netherlands across age groups and the years.

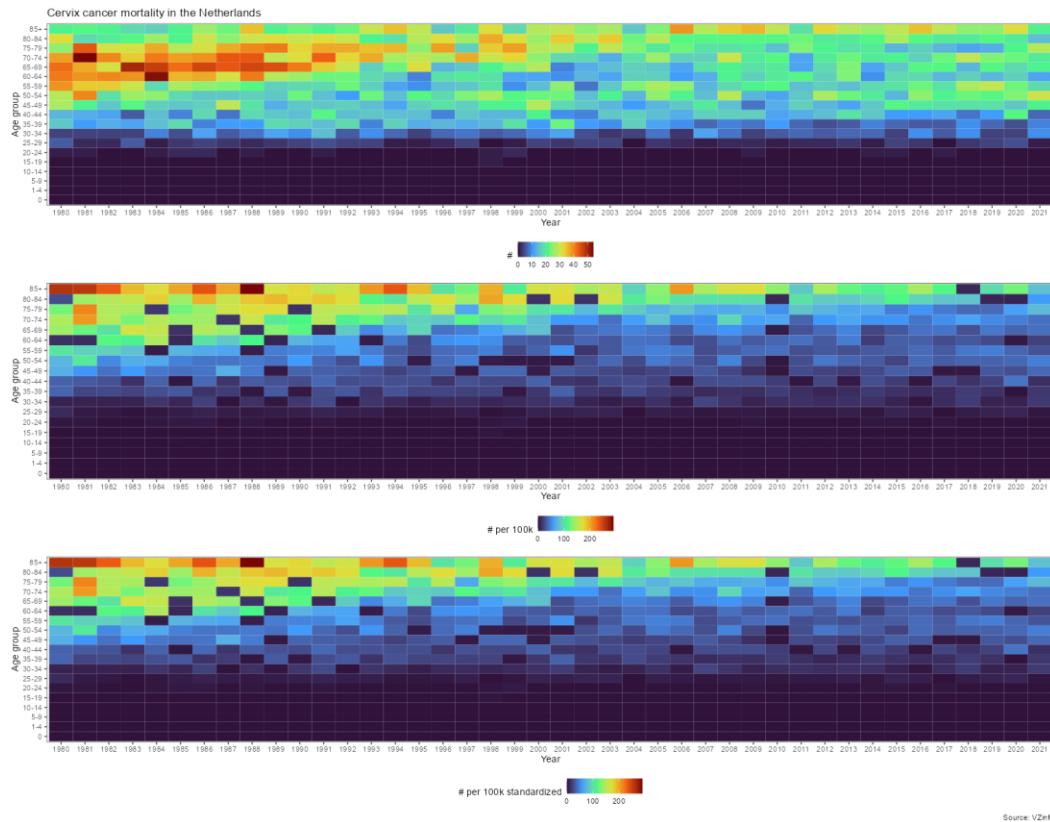


Figure 39. Cervical cancer mortality - standardized and unstandardized - across age groups and the years.

If we also look at the total cervical mortality and compare pre-Covid-19 to Covid-19 years, we see natural fluctuation but no big trend deviations (Figures 40 and 41).

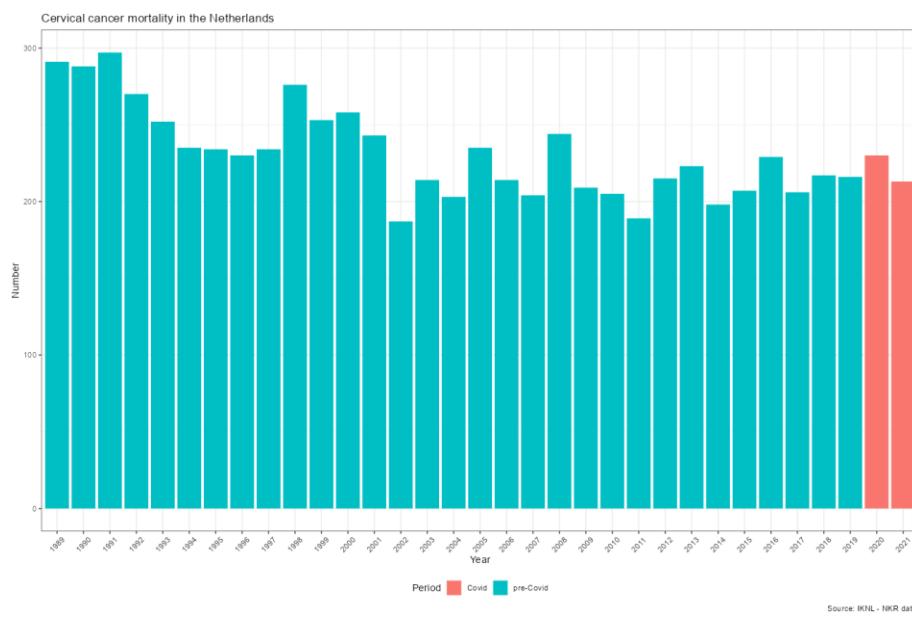


Figure 40. Total cervical cancer mortality during pre-Covid-19 and Covid-19 periods.



Figure 41. Total cervical cancer mortality during pre-Covid-19 and Covid-19 periods across ages.

Colon cancer

For colon cancer, we looked at the same sources as for breast and cervical. Also here, we found no large trend reversals as can be seen in Figures 42 – 48.

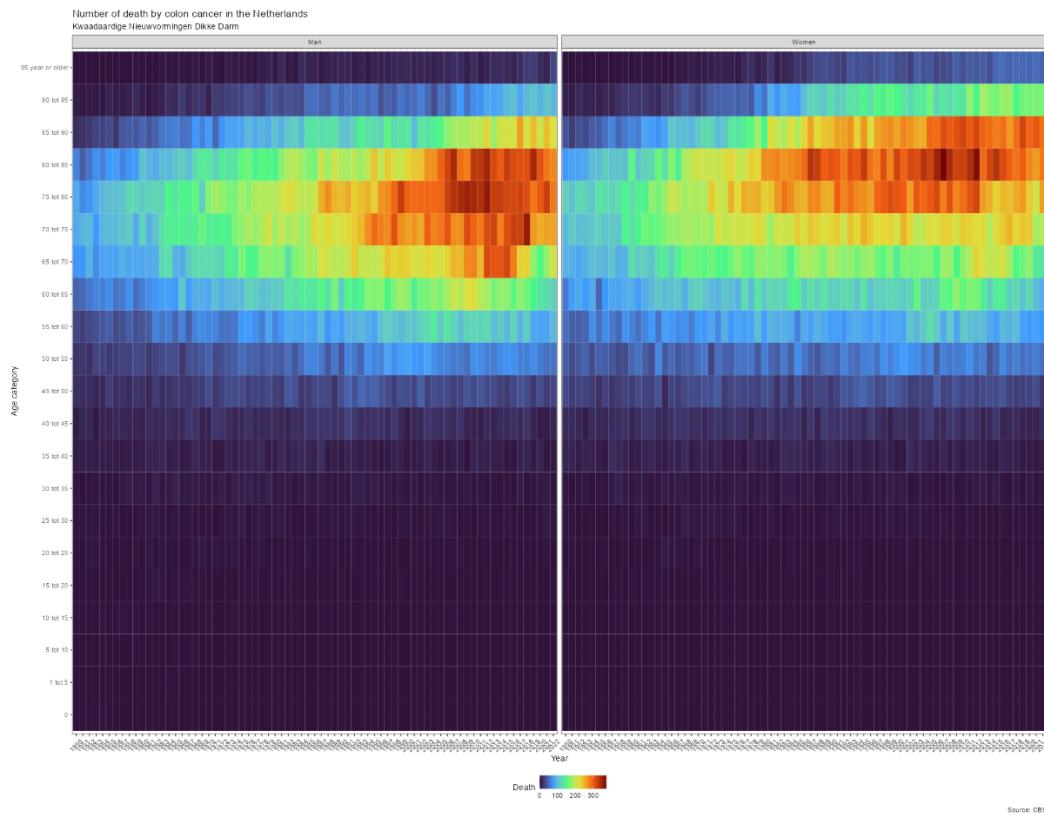


Figure 42. Number of colon cancer death in the Netherlands across years, age groups and sex.

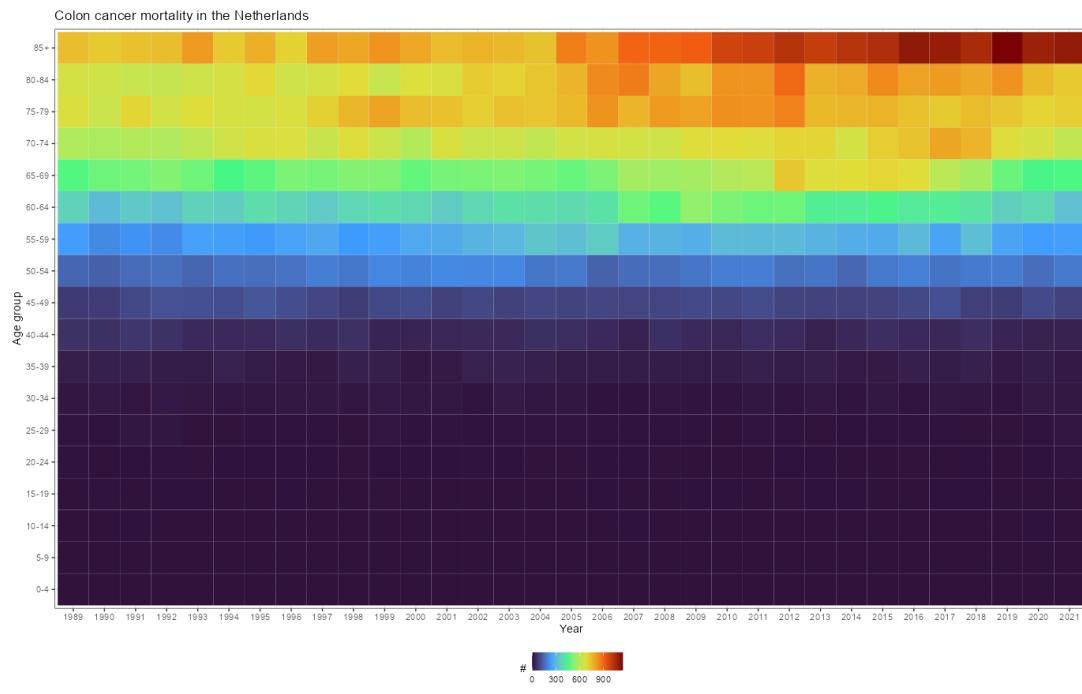


Figure 43. Colon cancer mortality numbers per age category coming from IKNL.

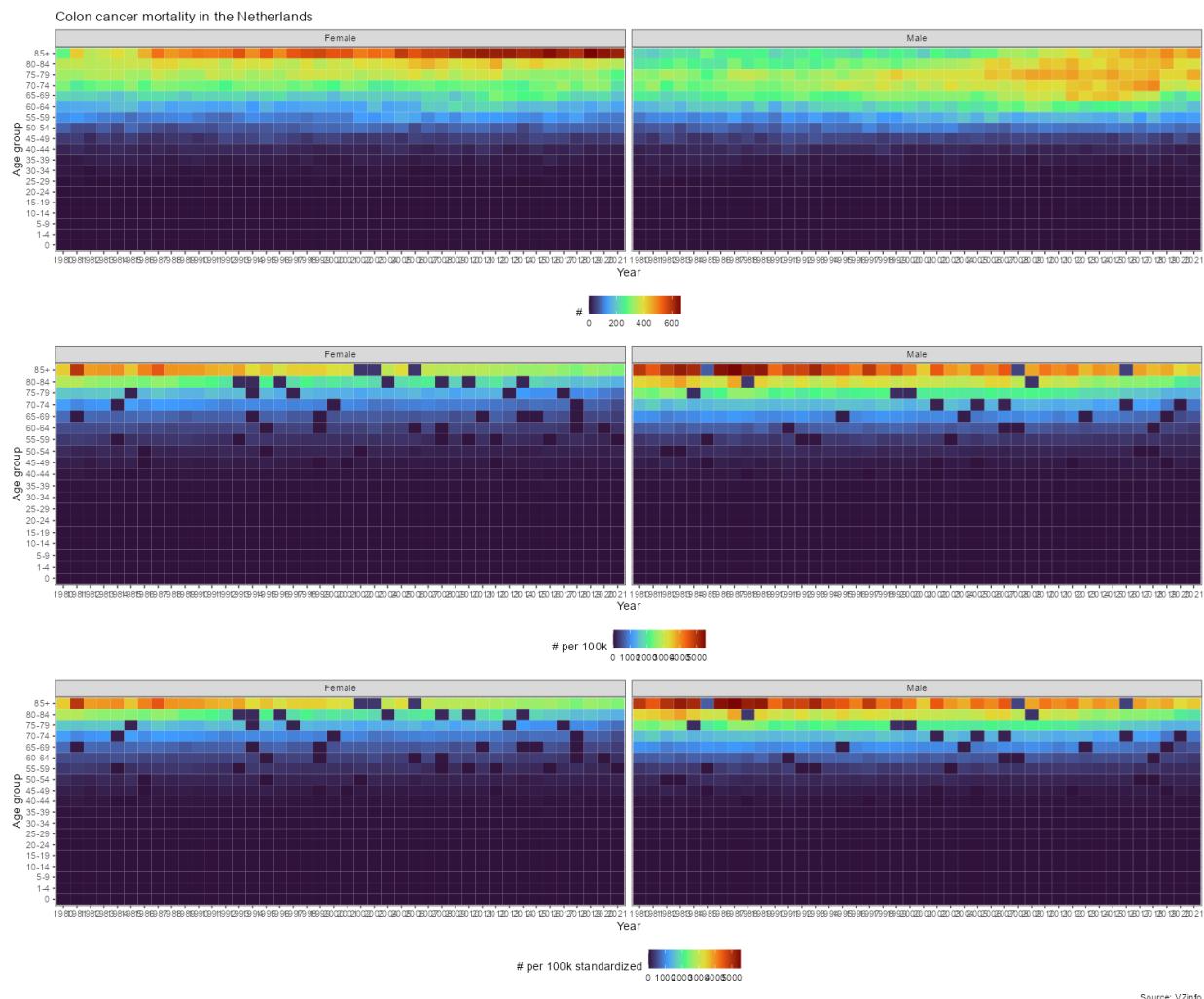


Figure 44. Colon cancer mortality numbers using raw data and standardized data from VZinfo across years and age groups.

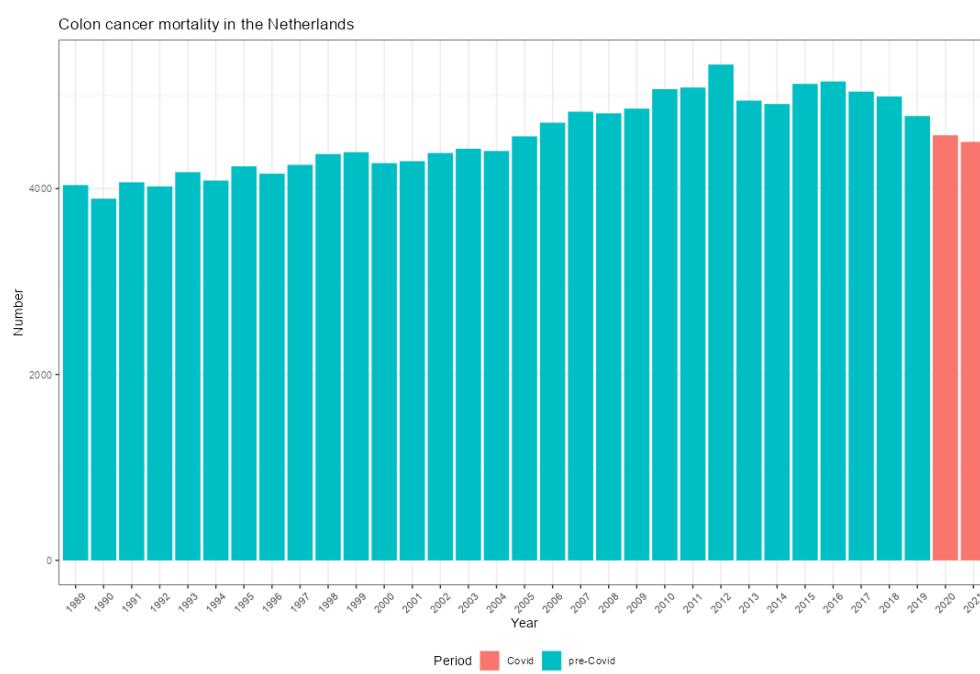


Figure 45. Colon cancer mortality compared pre-Covid-19 to Covid-19.

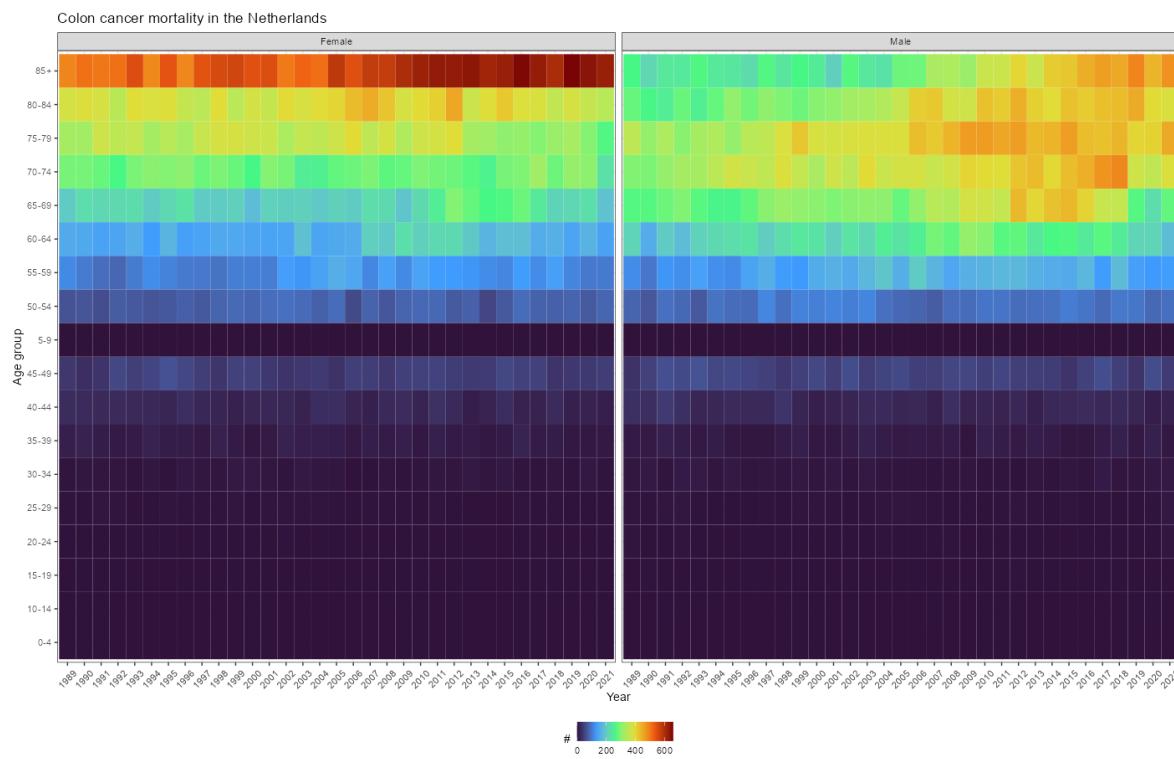
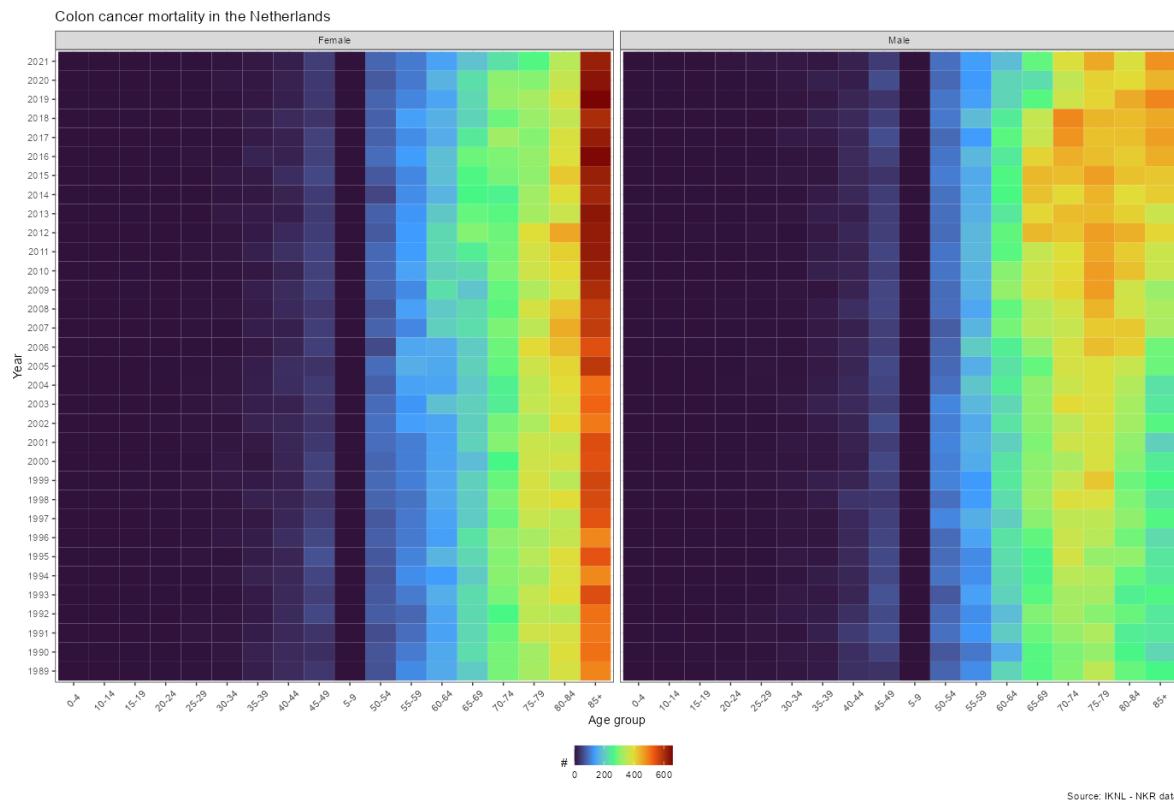


Figure 46. Colon cancer mortality numbers across the years per age category and per sex. Both graphs show the same data but have their axes inverted.

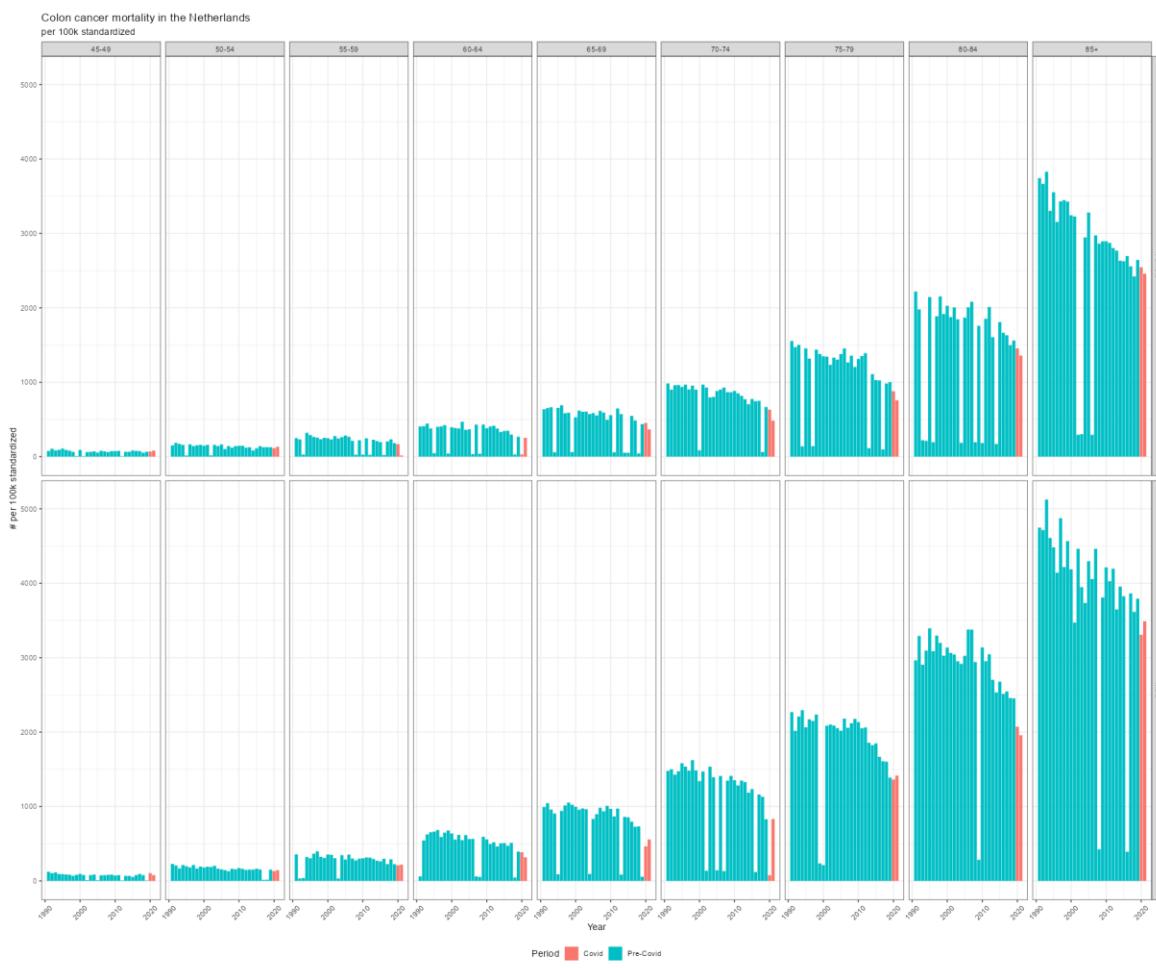
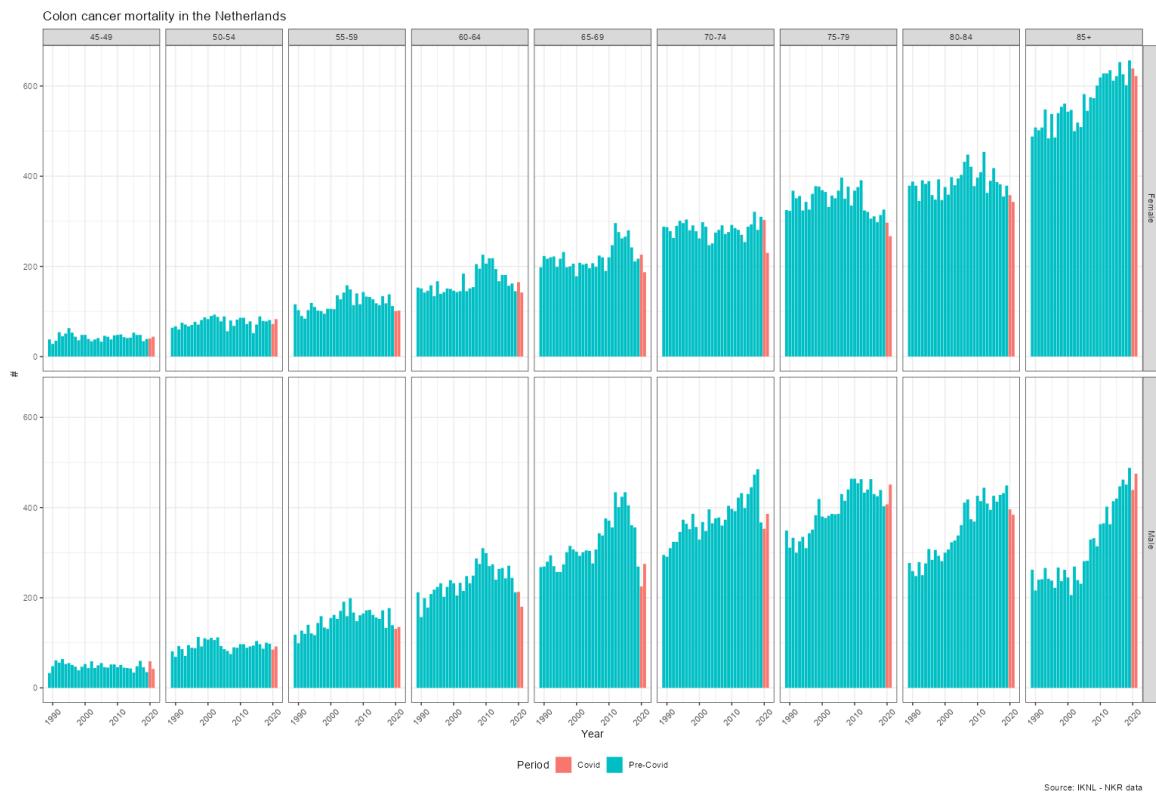


Figure 47. Colon cancer mortality per age group, year, and sex in which we compare pre-Covid-19 to Covid-19 years.



Figure 48. Same data as figure 47 but shown differently. Here, the colors show sex, not period.

Cancer incidence and prevalence

From cancer mortality we need to go to cancer incidence and prevalence. Since we are looking for the impact of delayed care, we would expect differences in these numbers because of the absence of screening. Once again, we can tackle multiple sources to get a view of the overall incidence and prevalence of cancer, and by specific cancer disease.

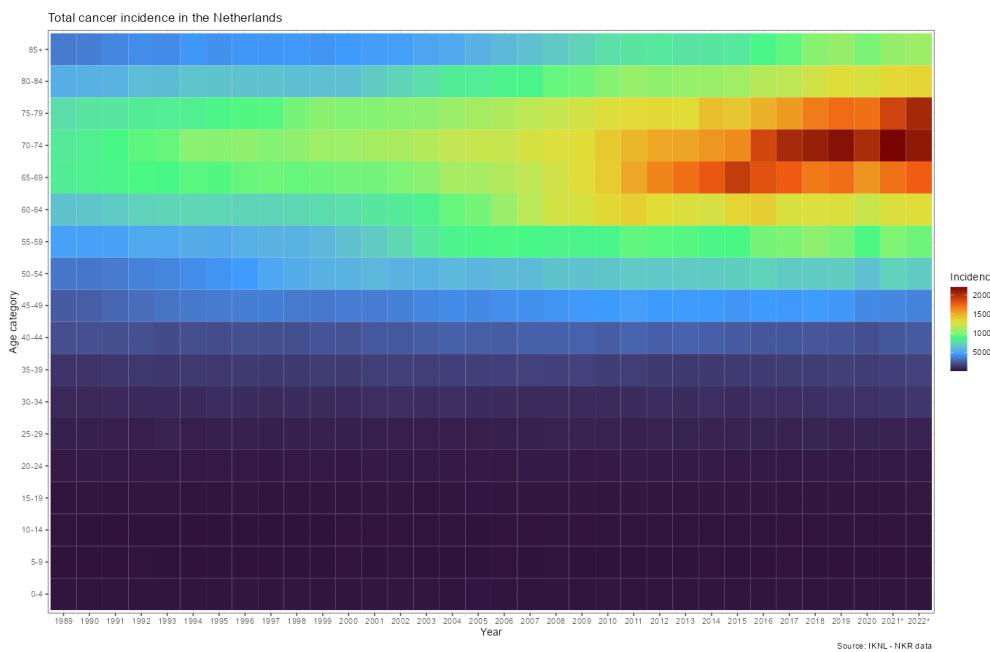


Figure 49. Total cancer incidence in the Netherlands per age category.

From Figure 49 it immediately becomes clear that cancer is rising. The diagonal line hints at a rise that is related to an increasing life expectancy and ageing population. No clear trend breach can be detected from the figure, except that the number of diagnoses in 2020 are lower than was to be expected.

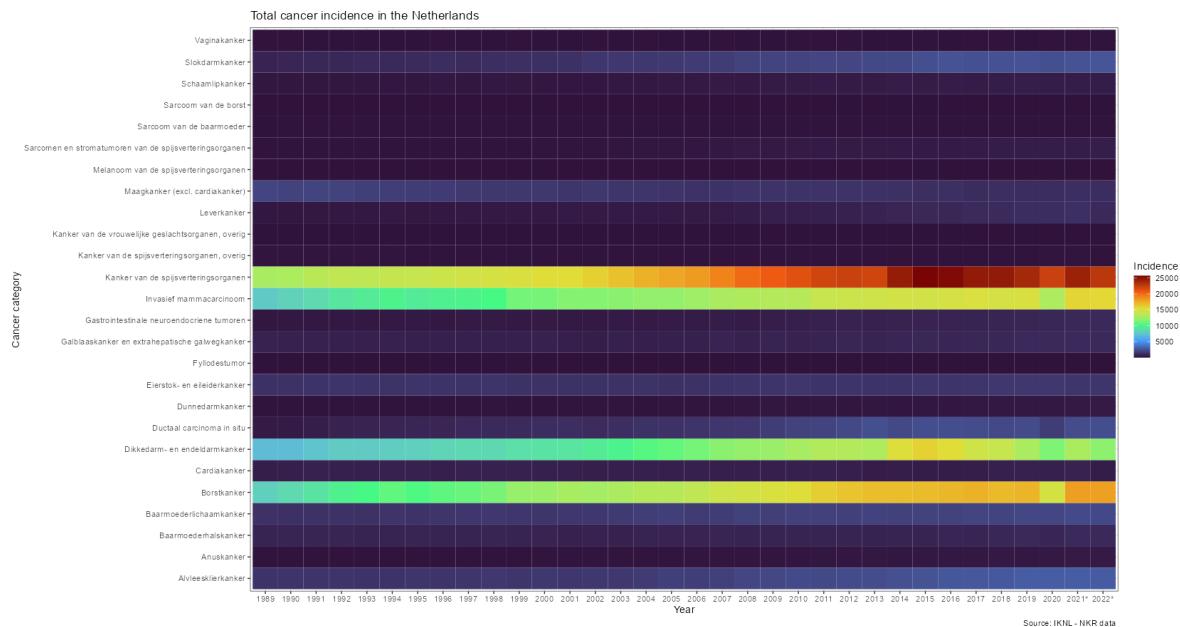


Figure 50. Total cancer incidence per major type of cancer across the years.

From Figure 50 it immediately becomes clear that breast cancer is the number one cancer in the Netherlands and that the raw incidence is rising. Second to breast cancer are cancers of the digestive tract, like colon cancer. There are also quite some differences in the crude total rate when looking at per province, but this is likely because of a difference in population. In the end, cancer has been rising overall, and 2020 was a clear trend reversal in incidence numbers (Figure 51).

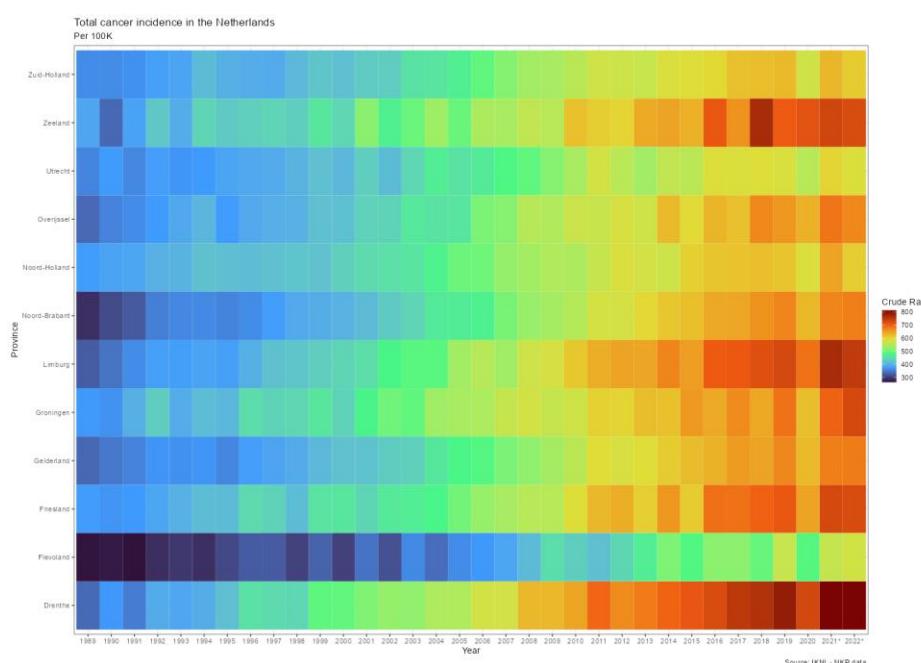


Figure 51. Total cancer incidence in the Netherlands (crude rate) per province and across the years.

Synthetic breast cancer data

An important part of our question is to determine how the various cancer stages have progressed over time. However, that data cannot be extracted from the NKR data. What we do have is synthetic breast cancer data which does have this information. However, before we can use that data we need to at least check if the descriptive patterns match the actual data. We will do this for variables for which we have data from multiple sources. Let's start by just comparing the incidence for breast cancer (invasive mamma) using both the NKR and the synthetic data from IKNL (Figure 52). It is clear to see that the synthetic data do not follow the NKR data. It is also unclear to us why IKNL does not provide tumor stage data for incidence.

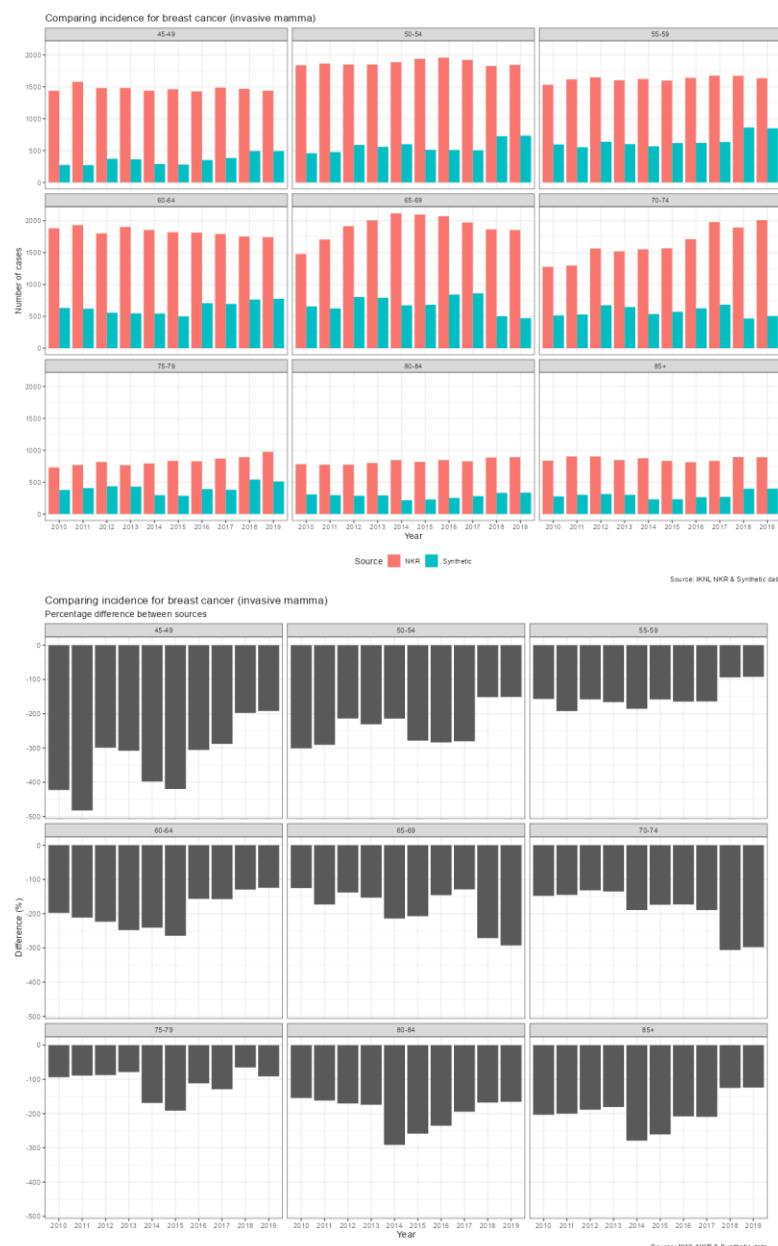


Figure 52. Comparing the number of breast cancer cases across NKR numbers and synthetic numbers per age category. The upper graph shows the real data, the lower graph the difference between both sources.

A second important variable that is not contained within the NKR data but is present in the synthetic data is the source of diagnosis. When looking at the vital status of each patient, given each year, and how they were originally diagnosed we can have a look at trends. For instance, do patients who come in from clinical-diagnostic research have a lower higher likelihood of dying in a given year. Of course, such an analysis is very limited due to not including many clinically important confounding variables. Nevertheless, we are looking for trends here, and especially trend reversals. Since the data is until 2019, we cannot see the 2020 and beyond effect on this data.

Figure 53 shows no trend, but some peculiarities especially in the number of patients alive via the clinical-diagnostic research route. However, since screening is not particularly mentioned, we cannot dive into this.



Figure 53. Vital status, inclusion year and diagnostic route.

We can now make the same graph, but instead of vital status we will look for tumor stage (Figure 54), tumor grade (Figure 55), and treatment received (Figure 56).

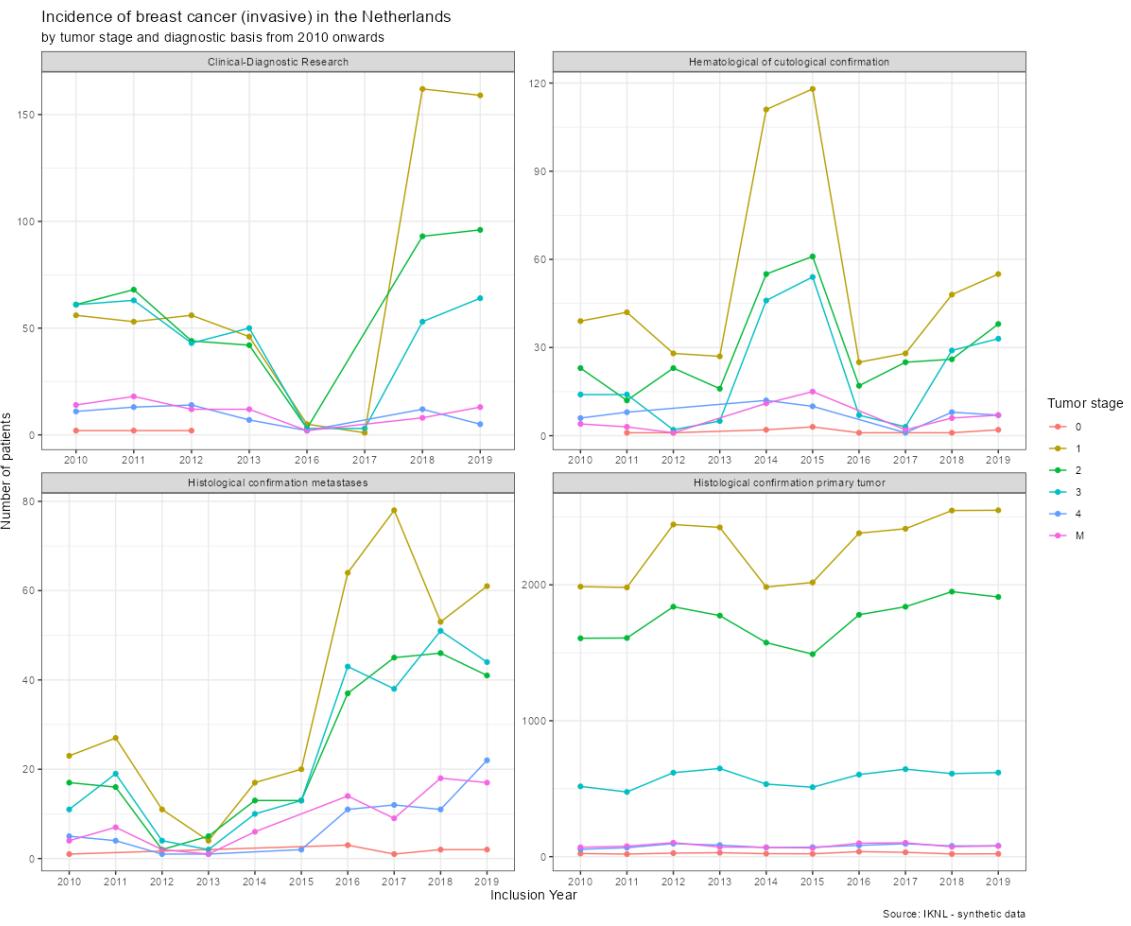


Figure 54. Incidence of breast cancer (invasive) per tumor stage and diagnostic basis from 2010 onwards.



Figure 55. Incidence of breast cancer (invasive) per tumor grade and diagnostic basis from 2010 onwards.

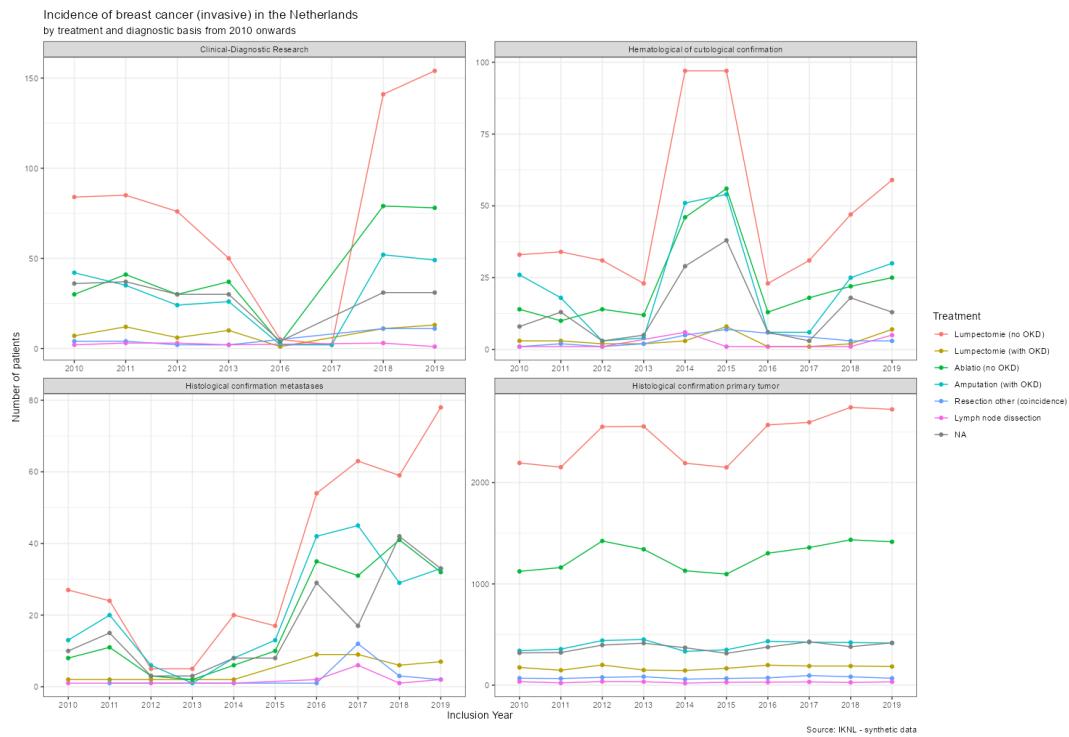


Figure 56. Incidence of breast cancer (invasive) per treatment received and diagnostic basis from 2010 onwards.

Following cancer incidence, it is time to look more closely at cancer prevalence. We will focus specifically on comparing the pre-Covid-19 years to the Covid-19 years to look for trend reversals. Immediately clear to see from Figure 57 is that 2021 affected cervical cancer and breast cancer, but not so much colon cancer. In fact, it seems as if the effect is truly minor indeed. An important caveat to make is that the 2021 and 2022 data are not yet fully confirmed here. In Figure 58 we also do not see a clear trend reversal for breast and cervical cancer, but we do see a trend reversal for colon cancer.

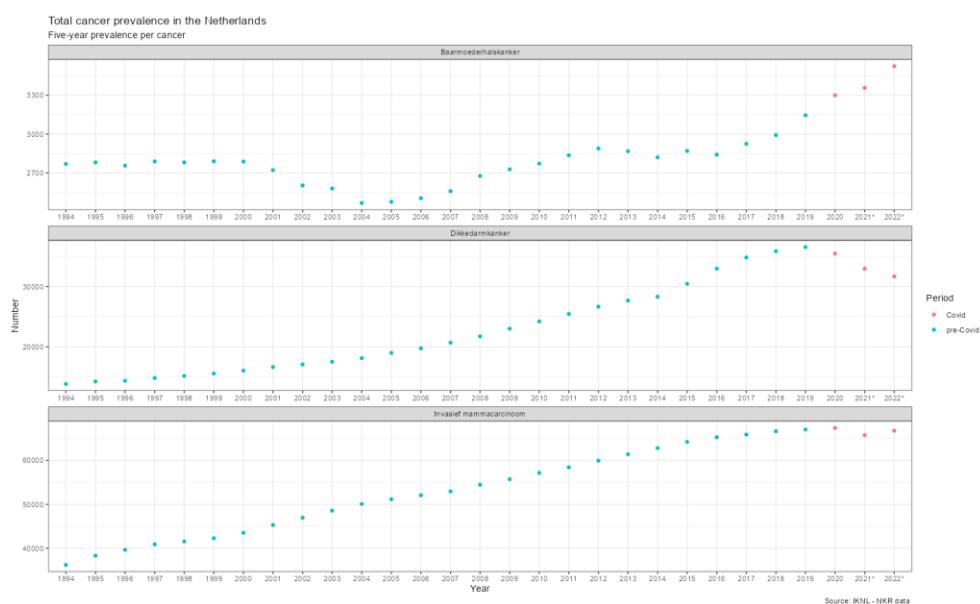


Figure 57. Total cancer prevalence for breast, colon and cervical cancer per year and colored by period.

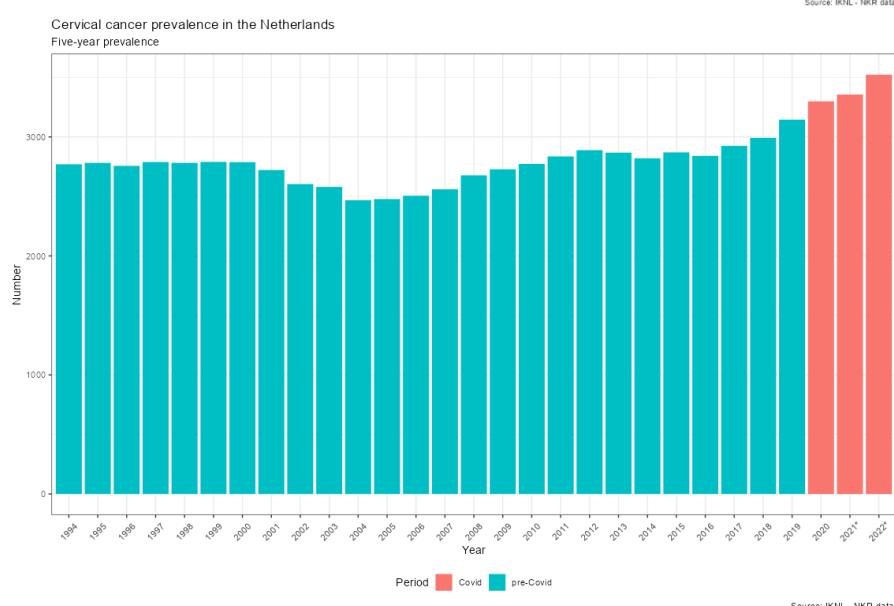
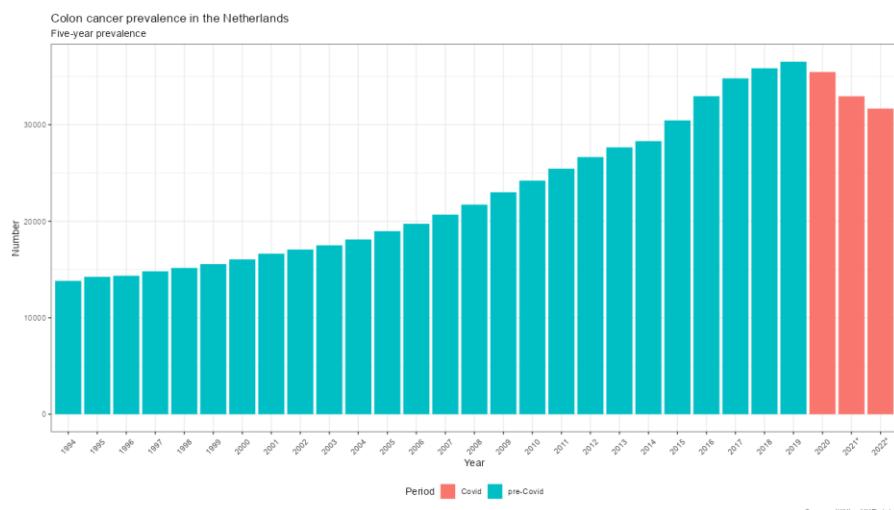
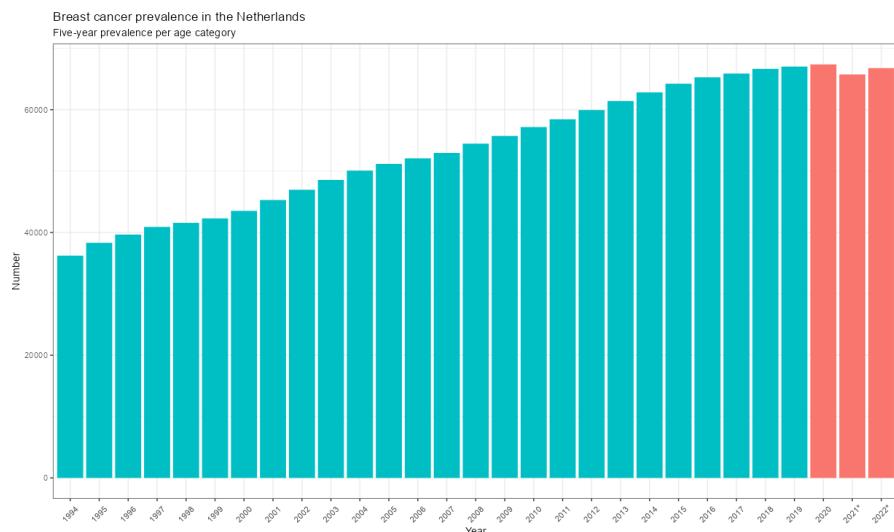


Figure 58. Total cancer prevalence for breast, colon and cervical cancer per year and colored by period. It is the same data as the figure before but shown using bar graphs instead of points.

Cancer screening

From cancer incidence and prevalence, we move towards cancer screening. We could already see in the previous pictures that both incidence and prevalence have shown somewhat of a decrease which is most likely due to the halt of screening methods. Let's look at screening data. The [OECD](#) has European data on screening rates so let's look at these first, since we know that Covid-19 and the response to Covid-19 was quite similar across countries. In figure 59 we see that screening rates are generally high in the Netherlands but dropped in 2020. From then on, the supply of data to the OECD also stopped for many countries.

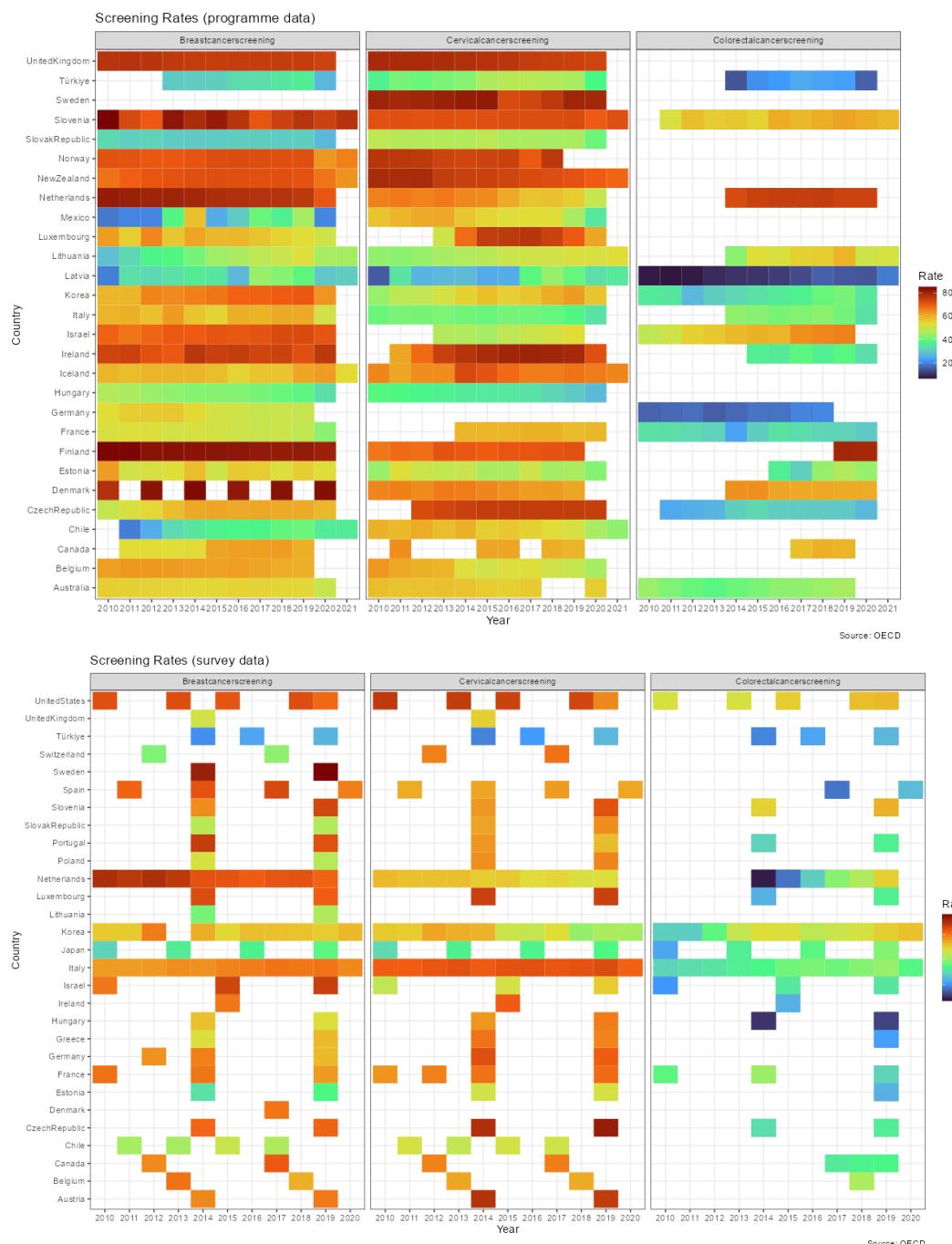


Figure 59. Screening rates using programme or survey data for European countries.

Looking at national data, coming from '[De Staat van Volksgezondheid en Zorg](#)' we can clearly see a dip in cancer screening rates for breast and cervix, but not colon (Figure 60). The cancer screening rates for breast have been decreasing for years it seems, but 2020 was a clear trend reversal. Age does not seem to have a particular effect on this.

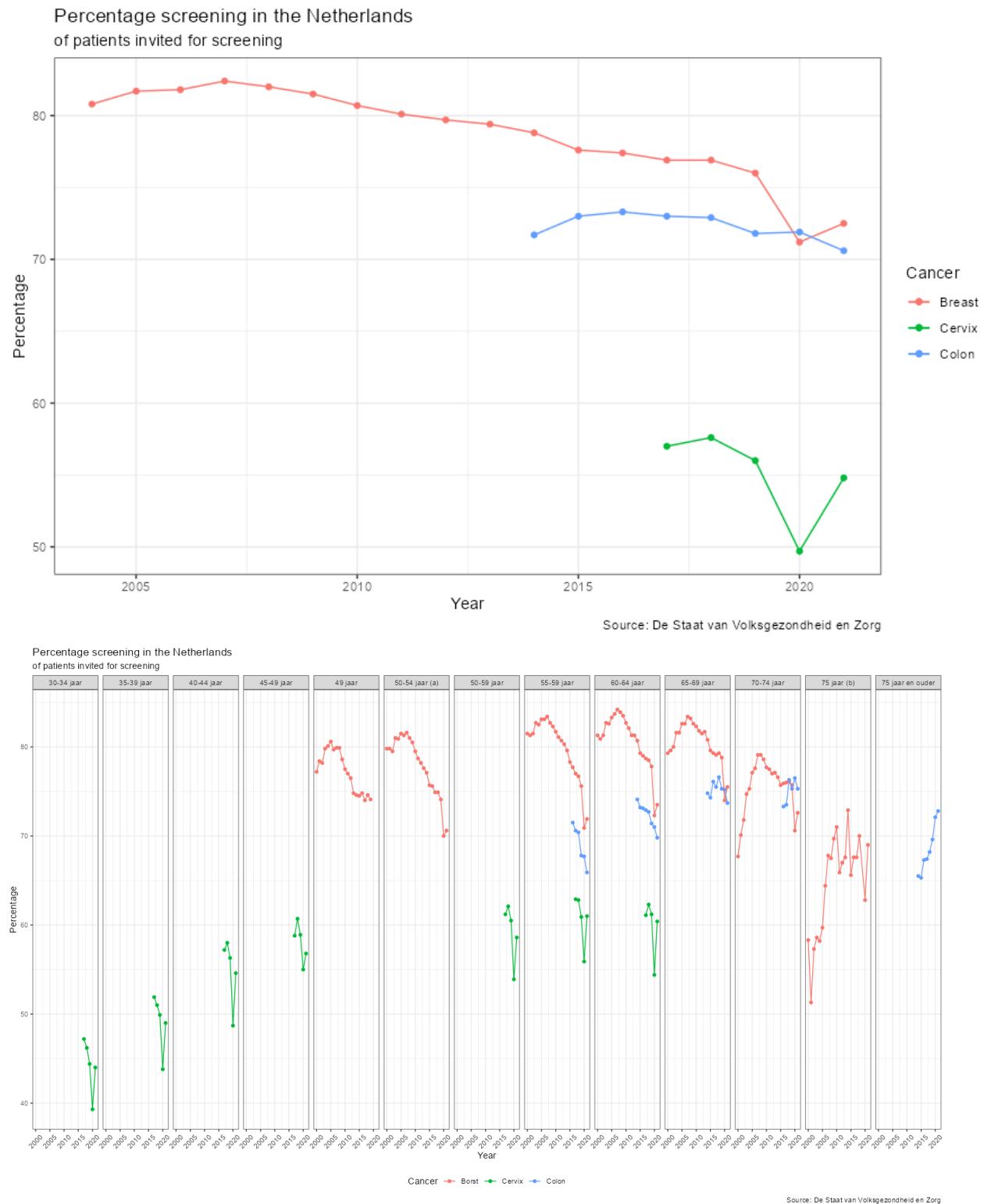


Figure 60. Dutch screening rates for breast, colon, and cervical cancer, split by age category.

Figure 61 shows the exact same data as Figure 60, but in heatmap form. Clear to see is how for breast and cervical cancer the year 2020 was a trend reversal for all ages. The screening for colon cancer was less effected. If we explicitly compare pre-Covid-19 years to Covid-19 years, we see more clearly the trend reversals (Figure 62).

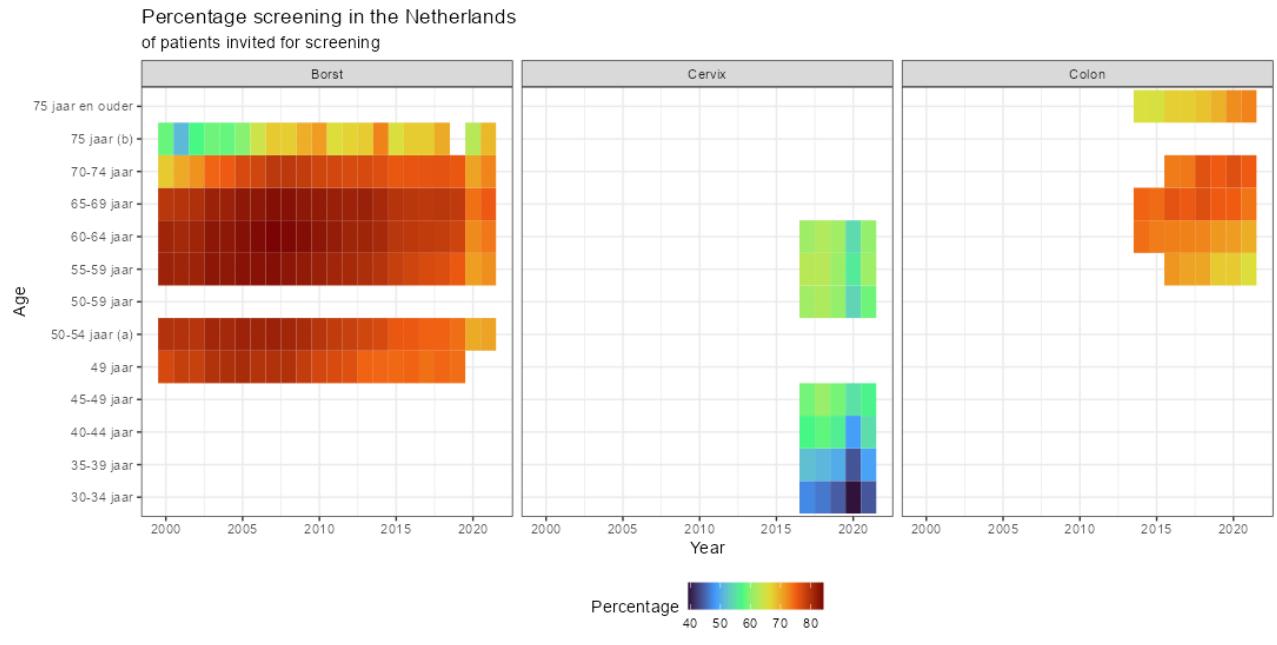


Figure 61. Percentage screening per age category for breast, cervical and colon cancer.

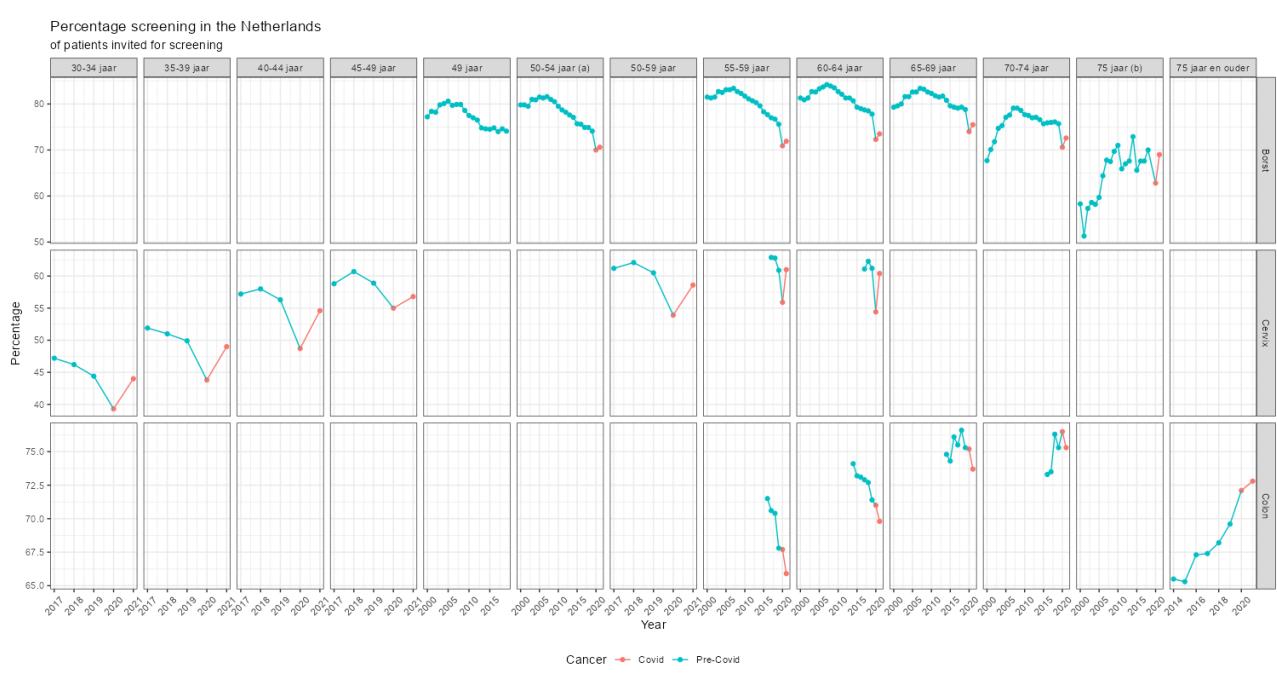


Figure 62. The percentage of patients invited for screening for breast, cervical, and colon cancer split by age category and colored by period.

Cancer survival

It is now time to investigate cancer survival. The first graph shows the data coming from the NKR dataset and has a particular focus on the number of patients that are alive following diagnosis for breast cancer in the Netherlands (Figure 63). These are five-year survival rates per TNM stage from 2010 onwards. The picture clearly shows that higher stages have lower rate to begin with, but also that time since diagnosis plays an important role. Most patients diagnosed are stage I and II patients. We see no trend reversals which could happen in year 0 and 1.

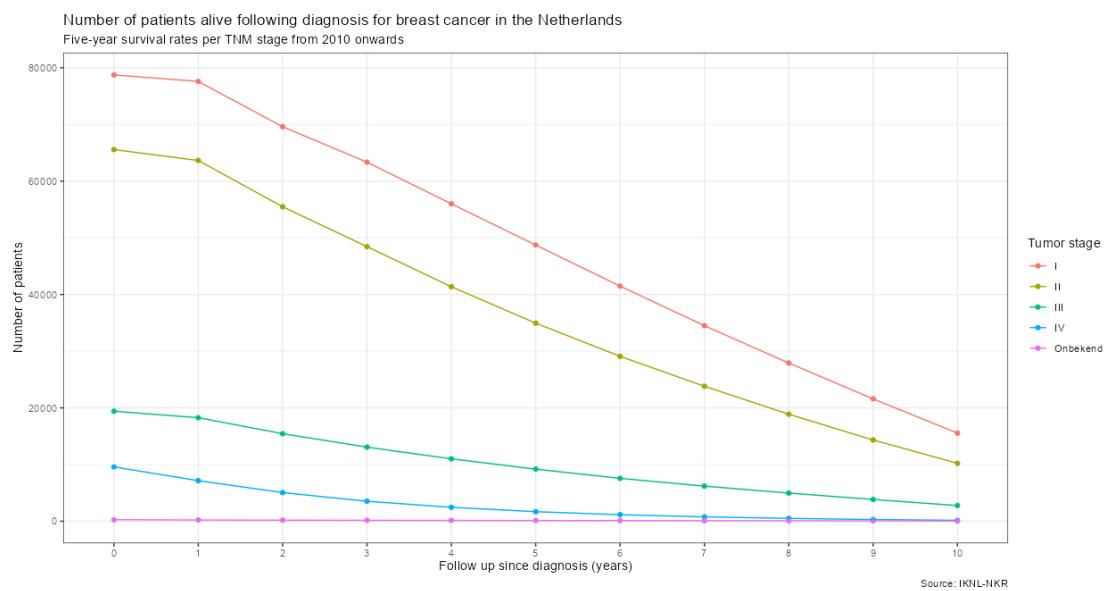


Figure 63. Number of patients alive following diagnosis for breast cancer in the Netherlands per tumor stage using five-year survival rates.

Perhaps a better graph to show the relationship between time since diagnosis, survival, and tumor stage is to plot the number of patients alive since diagnosis as both raw numbers and percentages (Figure 64). Here, it becomes clear that a higher tumor stage is less rare and has a worse survival rate.

We can also use the [synthetic breast cancer dataset from IKNL](#) to build survival curves per tumor stage, but the curves from a basis Kaplan-Meier curve look strange to be honest (Figure 65) It is unclear how effective this dataset truly is considering the high number of censorings going on and the inability of the dataset to distinguish between stages which at the very least should happen.

If we look more closely at cervical (Figures 66 and 67) and colon cancer (Figures 68 and 69), we see the exact same patterns as for breast cancer: lower stages are rarer and have a worse prognosis than a higher stage. Also here, we see no clear trend reversals for years 0 and 1, which could be expected given that the data is coming from 2022 and we are looking into 2020 and 2021. Nevertheless, we see no trend reversals.

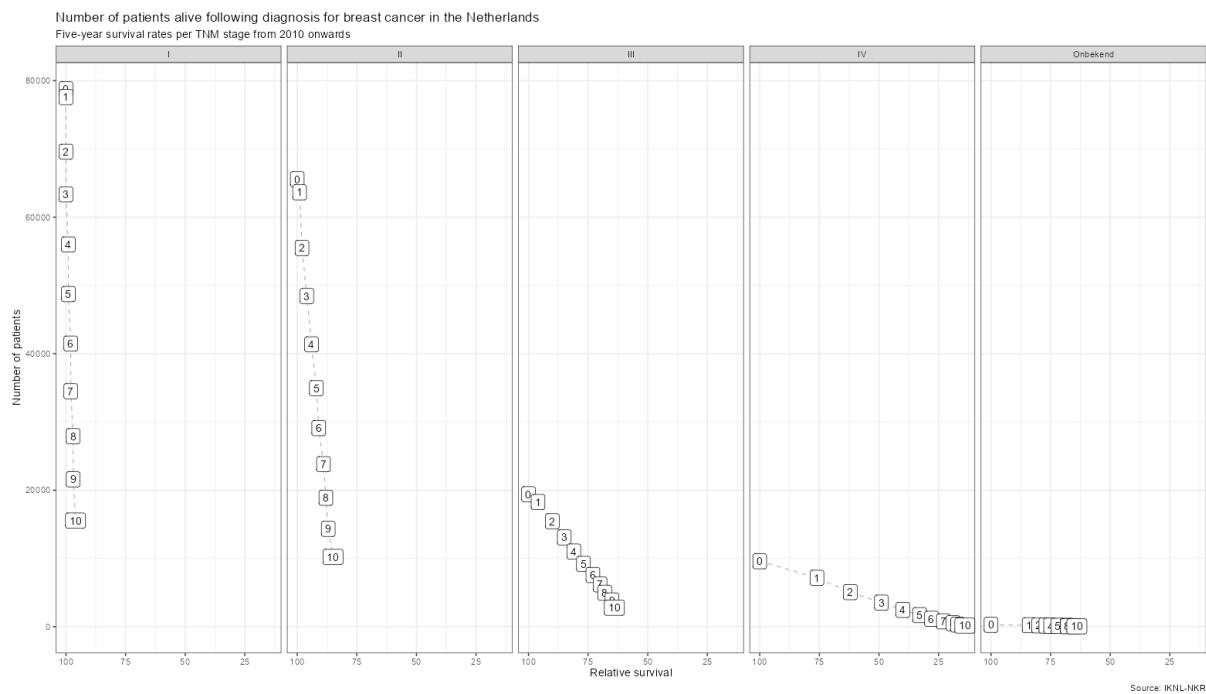


Figure 64. Number of patients alive following diagnosis for breast cancer in the Netherlands based on TNM stage and five-year survival rates.

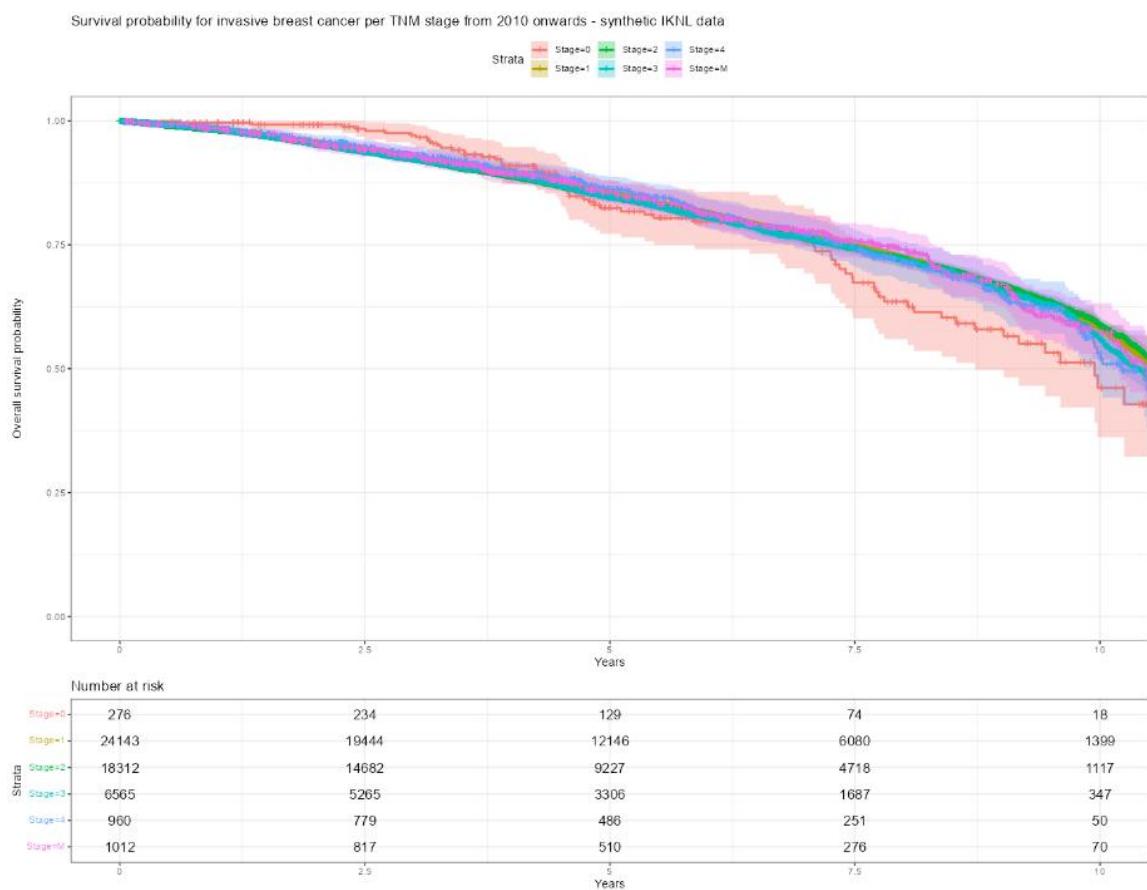


Figure 65. Kaplan-Meier curve showing the number at risk as a function of time since diagnosis and tumor stage for breast cancer patients using the IKNL synthetic data.

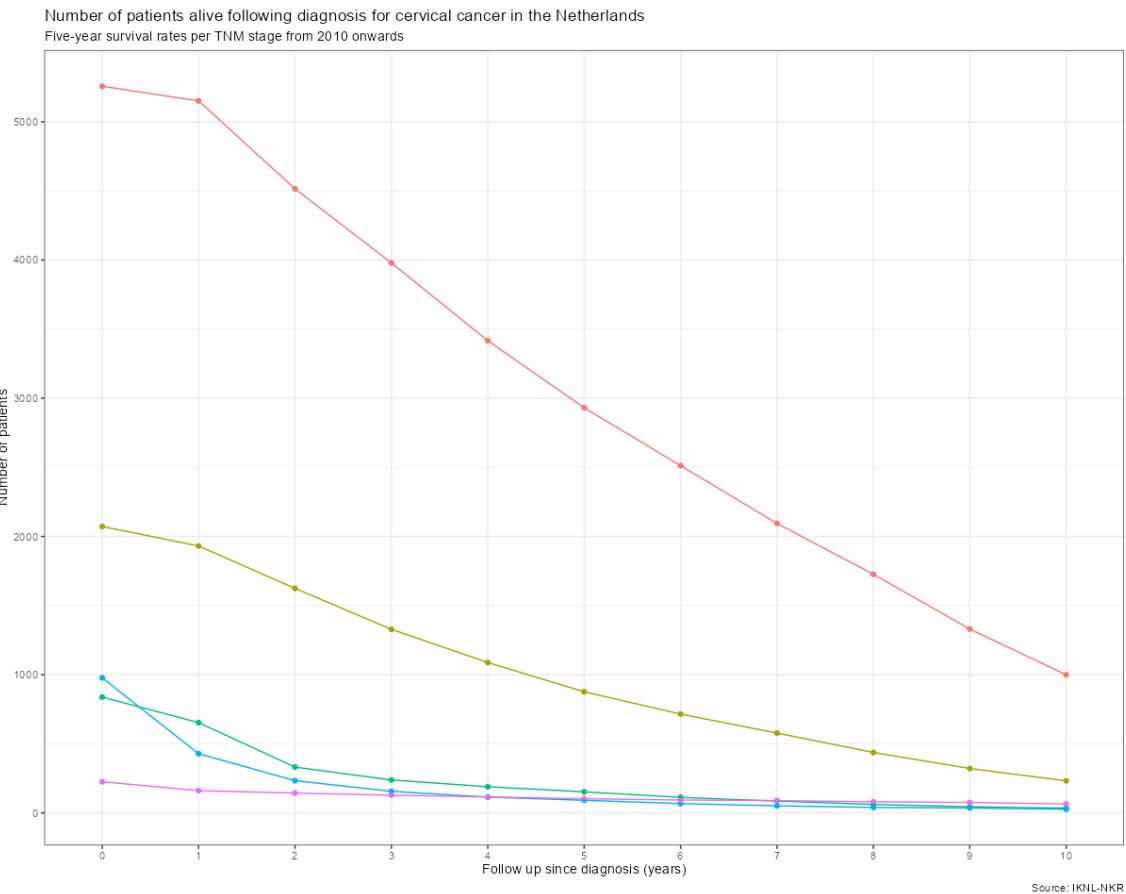


Figure 66. Number of patients alive following diagnosis for cervical cancer using five-year survival rates and TNM stage.

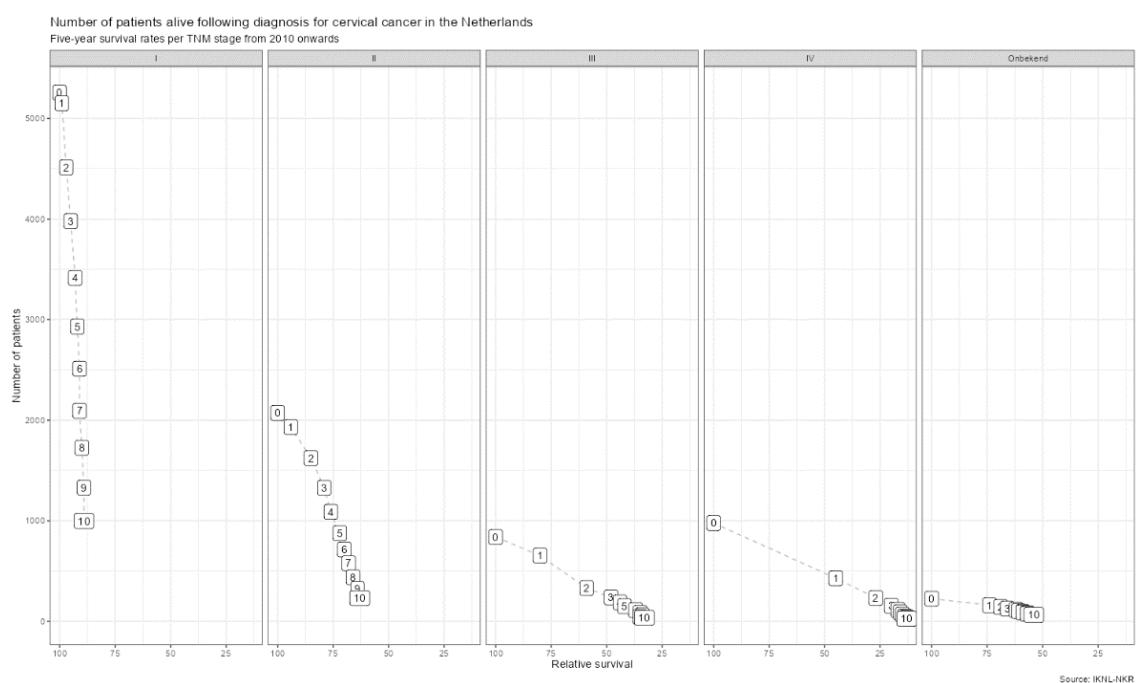


Figure 67. Number of patients alive following diagnosis for cervical cancer using five-year survival rates and TNM stage.

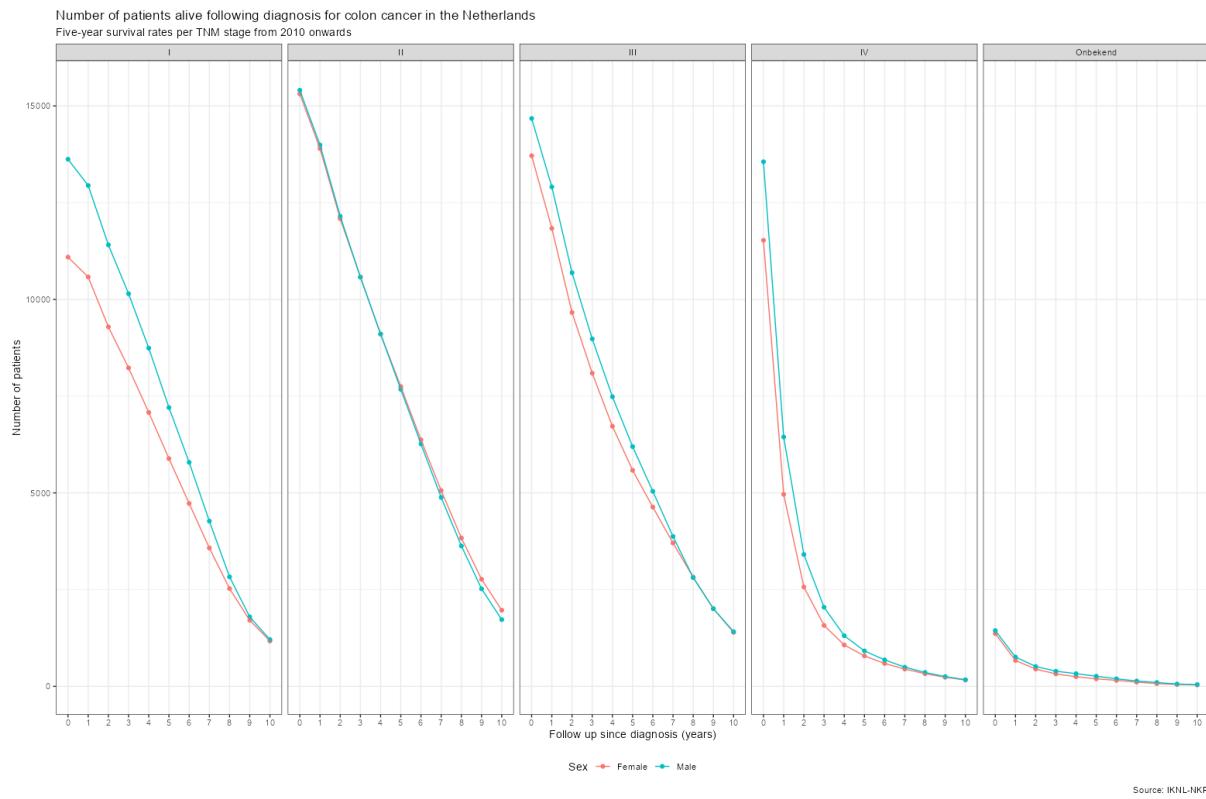


Figure 68. Number of patients alive following diagnosis for colon cancer using five-year survival rates, TNM stage, and sex.

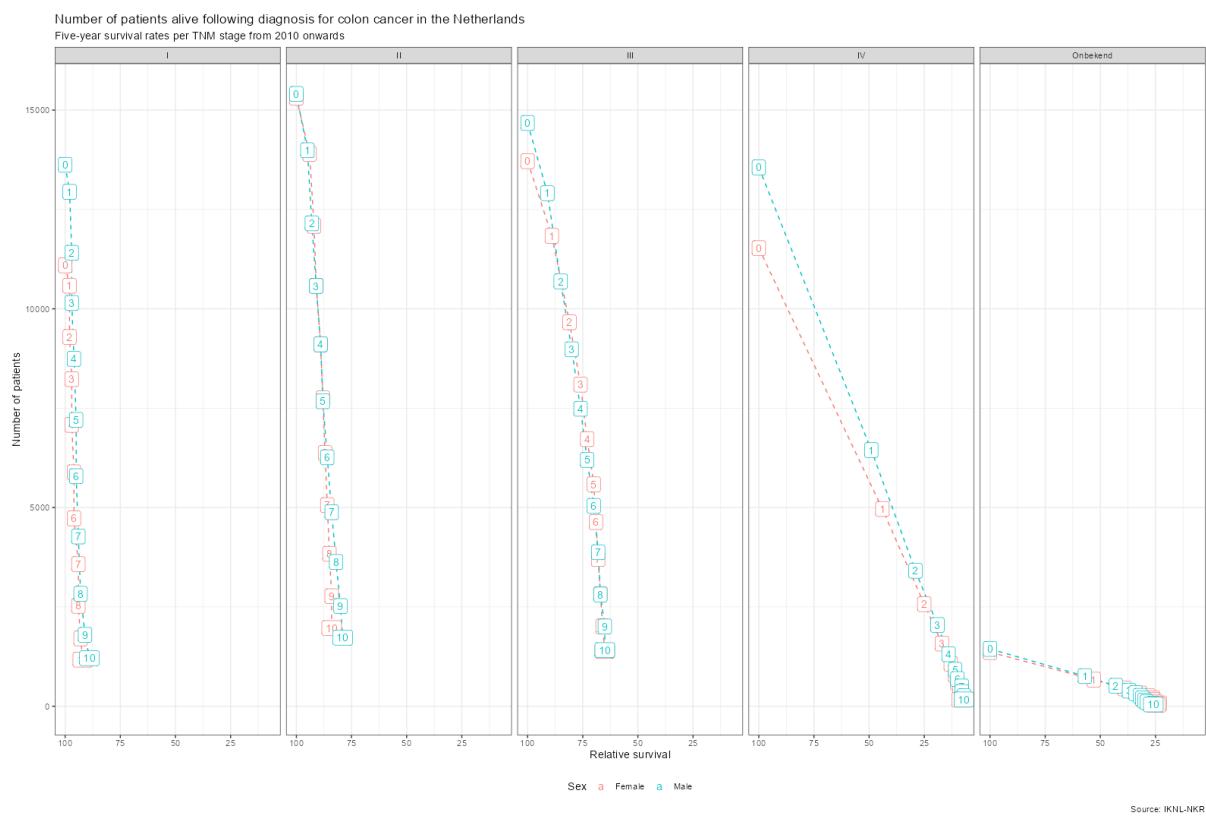


Figure 69. Number of patients alive following diagnosis for colon cancer using five-year survival rates, TNM stage, and sex.

Healthcare expenditure

The last part is the healthcare expenditure. First, we will look at expenditure using [VEKTIS](#) data before moving into the [OpenDIS](#) data. Both Figures 70 and 71 show that health-care expenditure dropped in 2020 compared to previous years of steady increase. Figure 71 shows that the expenditure is higher for males and females.

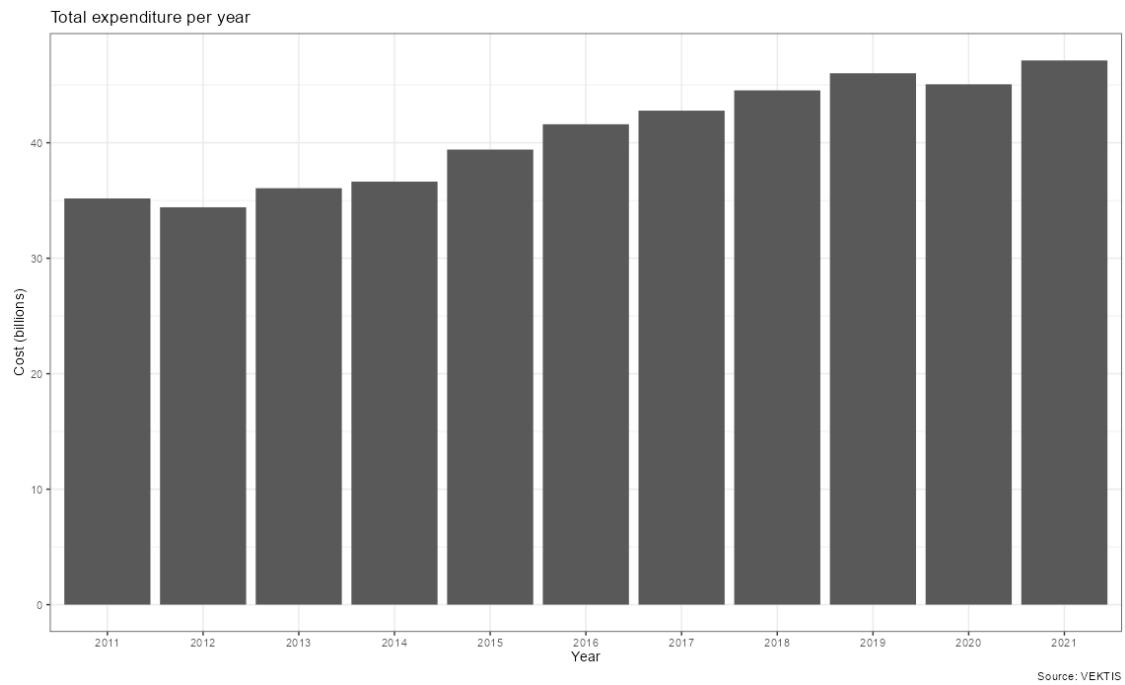


Figure 70. Total expenditure per year.

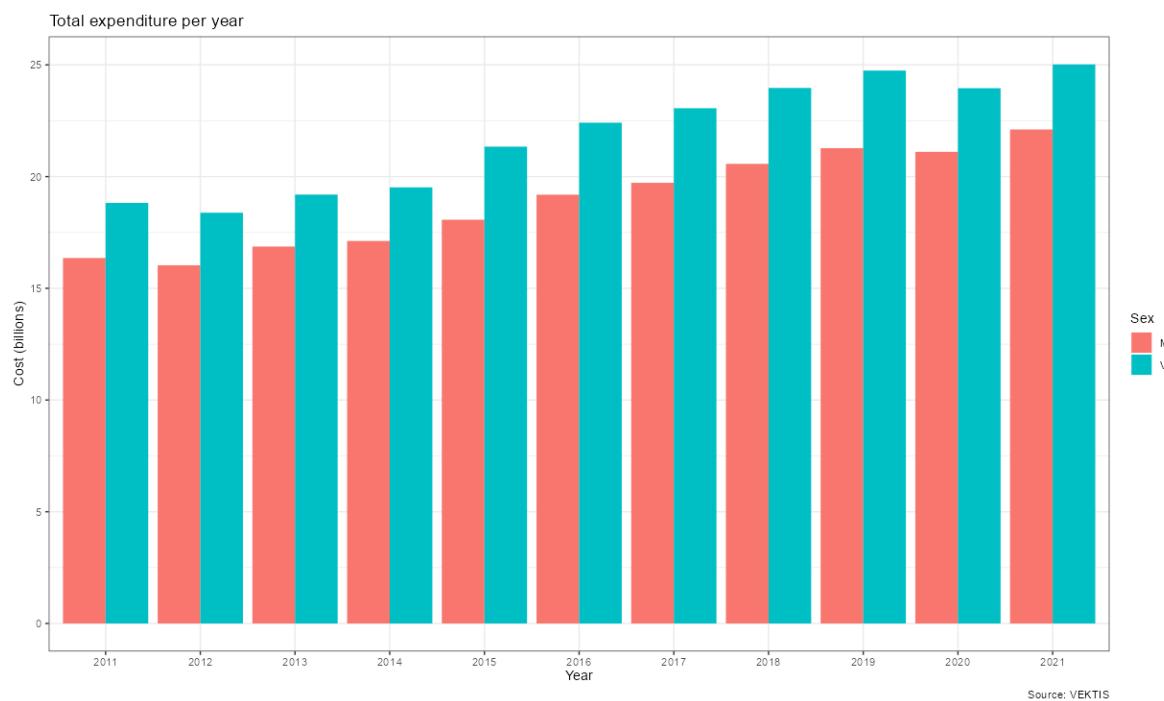


Figure 71. Total expenditure per year per sex.

The VEKTIS dataset offers data on healthcare expenditure on the level of a three-digit postal code. We wanted to see how much variation there is, per postal code, so we plotted the average expenditure per postal code, year, and sex (Figure 72). In addition, we focused on a single random postal code to look more closely at specialist medical care (Figure 73) and did another random search for 10 postal codes on specialist medical care (Figure 74). All three images show that there is quite a big variation between postal codes.

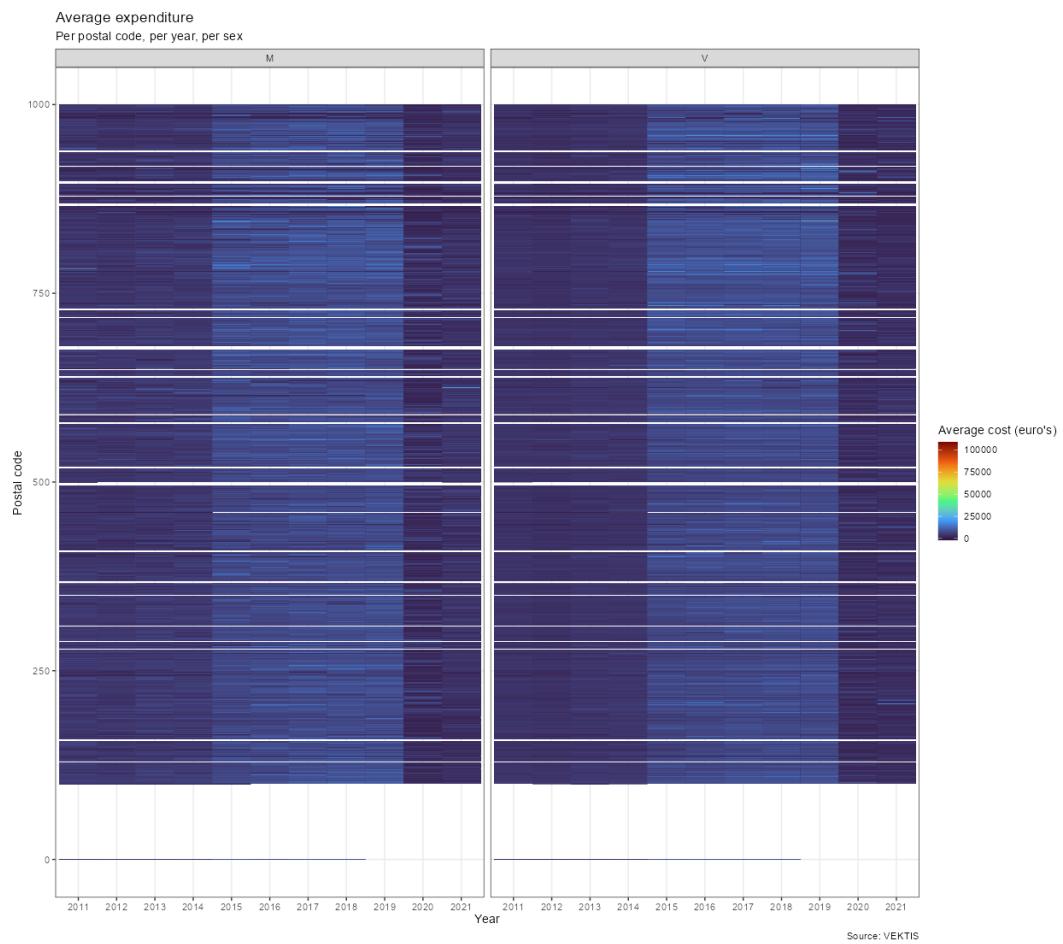


Figure 72. Average expenditure per postal code, per year and per gender.

From the above picture it becomes clear to see that we cannot just talk about the average postal codes. Therefore, it is perhaps best to look at the variation more closely using boxplots. Figure 75 shows per sex and year the total expenditure. Clear to see is that expenditure follows a very skewed distribution, with the median expenditure per year being less than one million, with outliers easily exceeding two to five million. A log transformation make confirms the skewness of the data (Figure 76). From the next three pictures (Figures 77 - 79), we can clearly see that age plays a major role in health-care expenditure, but that across postal codes there is an enormous variation. Hence, VEKTIS data is nice to have when aiming for a birds-eye view but does not offer additional information without being paired with clinical data.

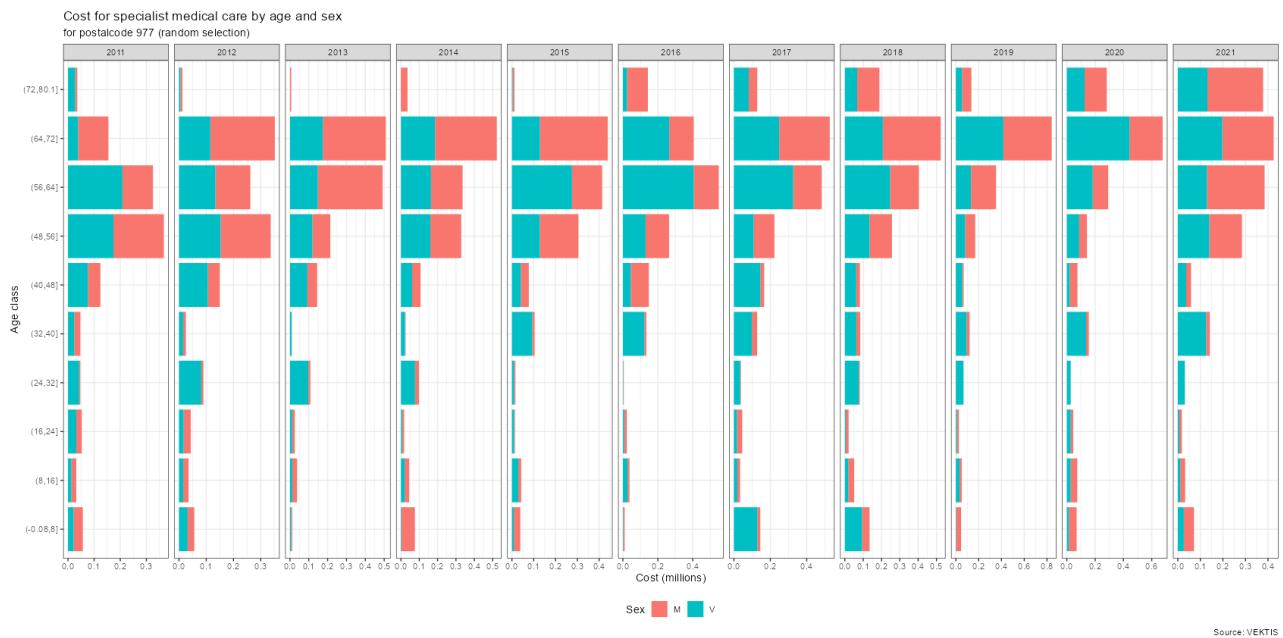


Figure 73. Cost for specialist medical by year, age and gender for postal code 977.

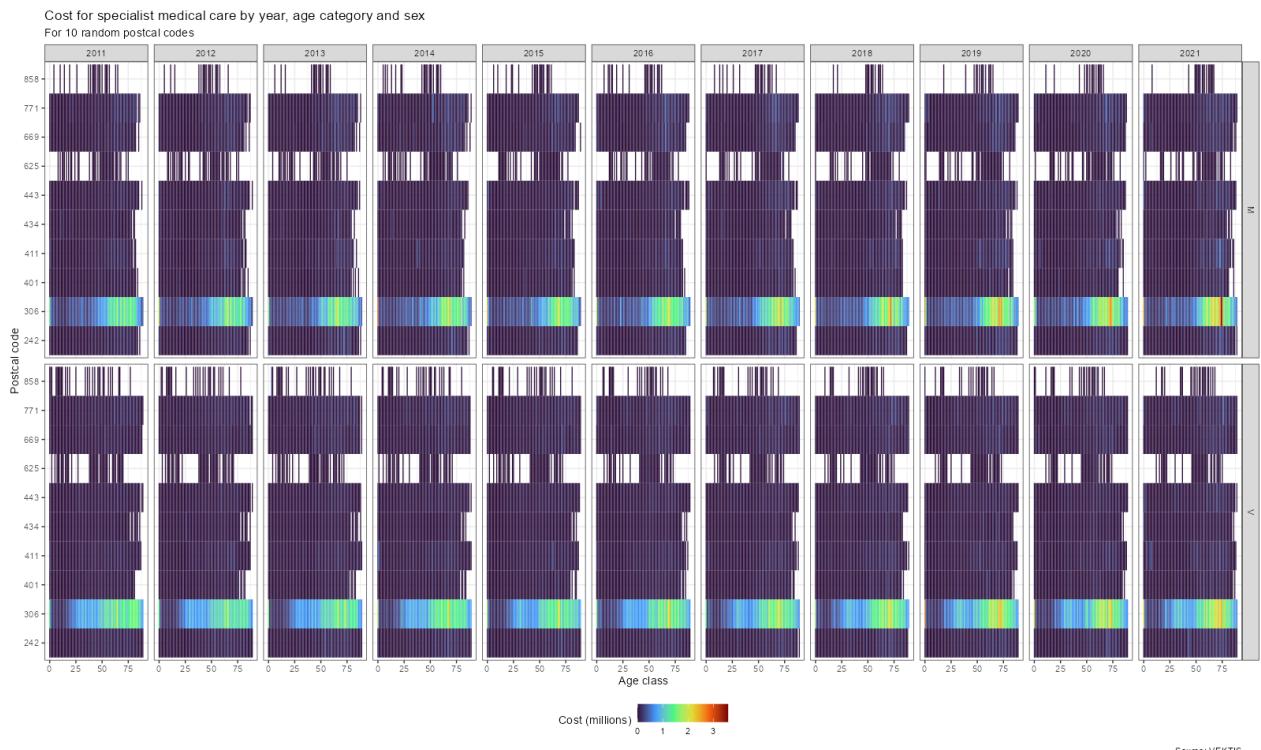


Figure 74. Cost for specialist medical by year, age category and gender for 10 random postal codes.

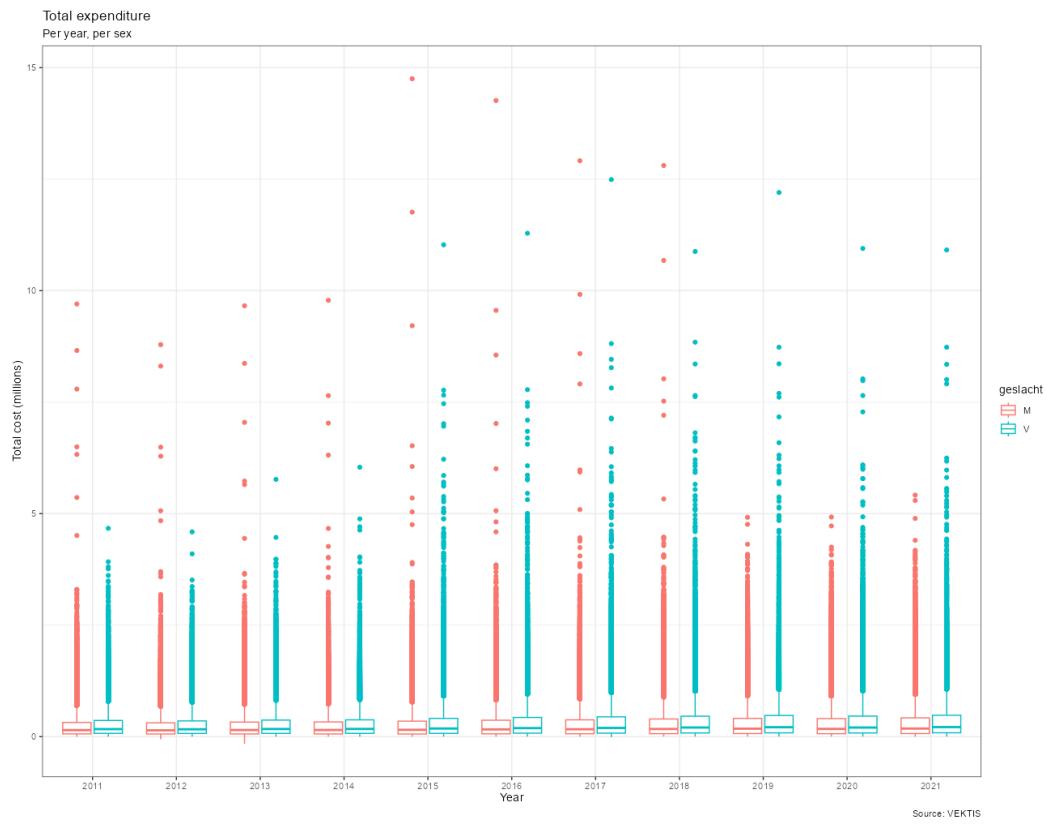


Figure 75. Total expenditure per year and per gender for each postal code.

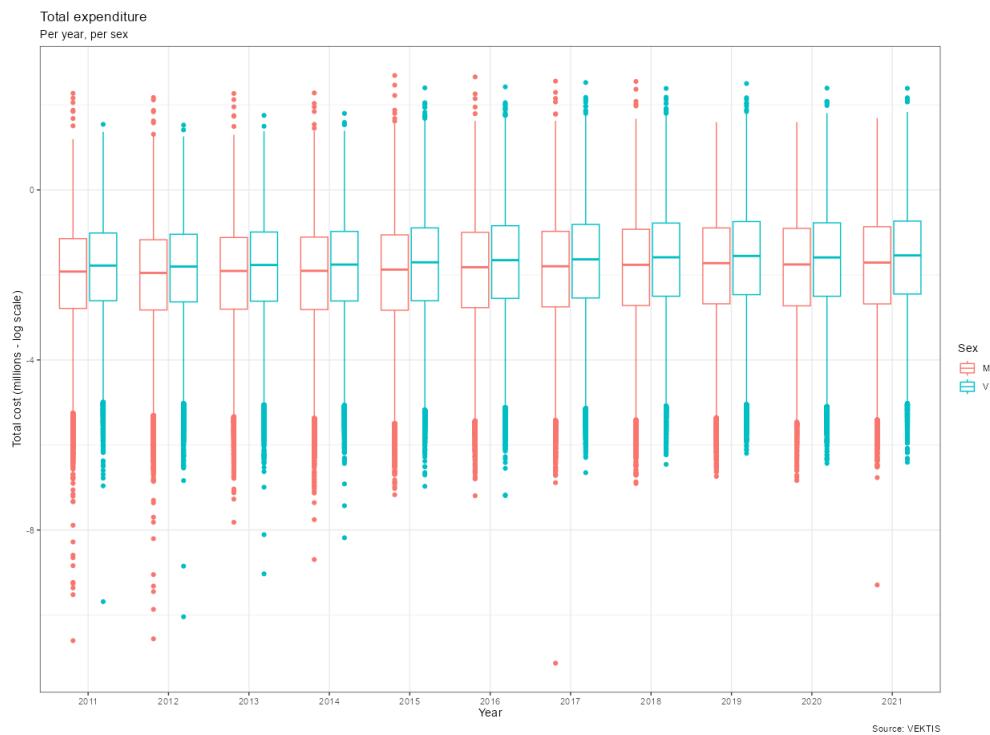


Figure 76. Total expenditure per year and per gender for each postal code (log scale).

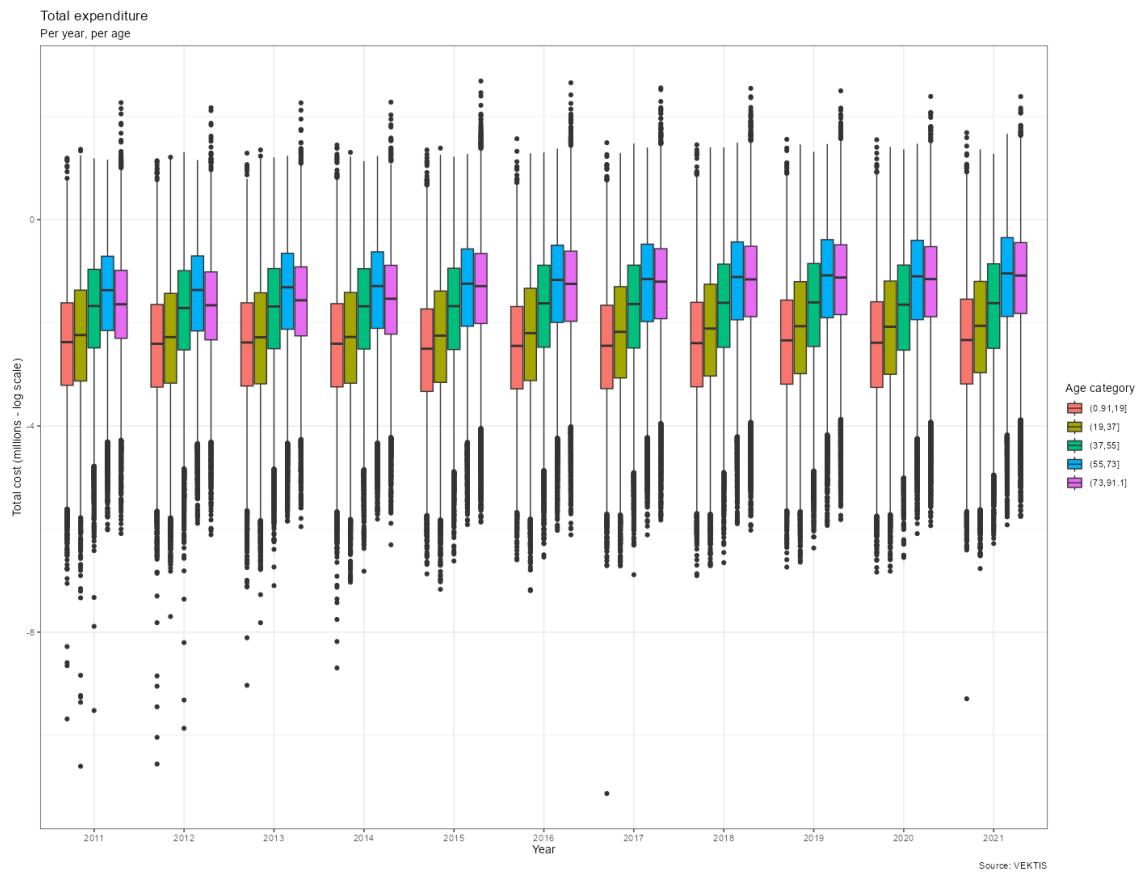


Figure 77. Total expenditure per year and per age category for each postal code (log scale).

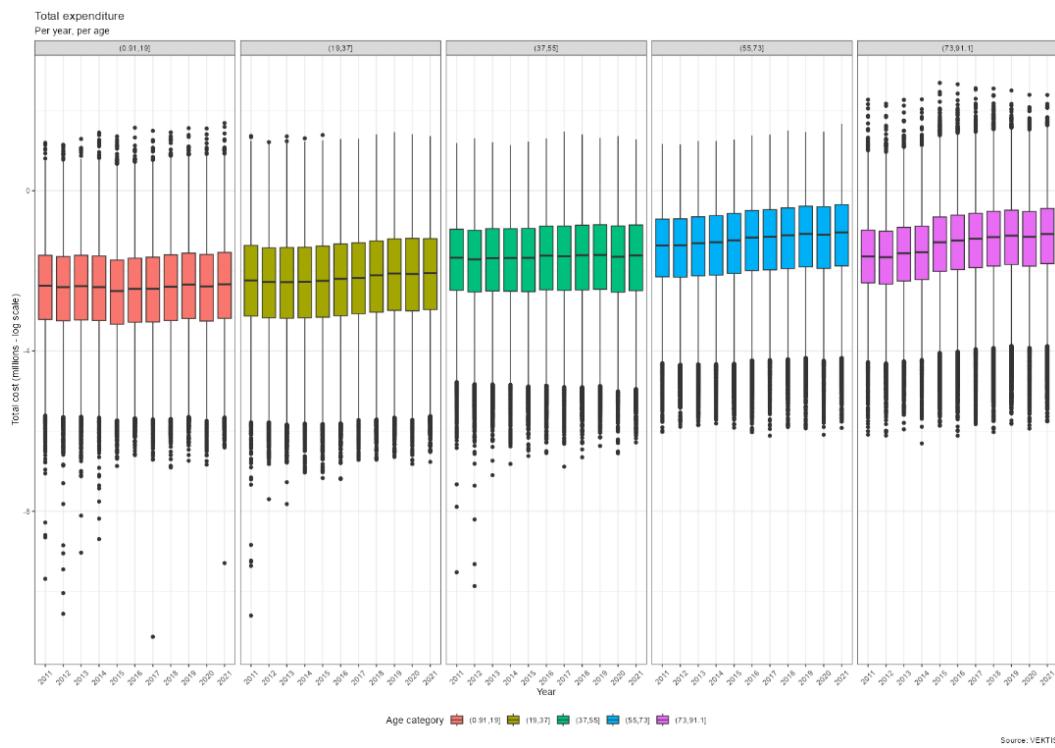


Figure 78. Total expenditure per year and per age category for each postal code (log scale).

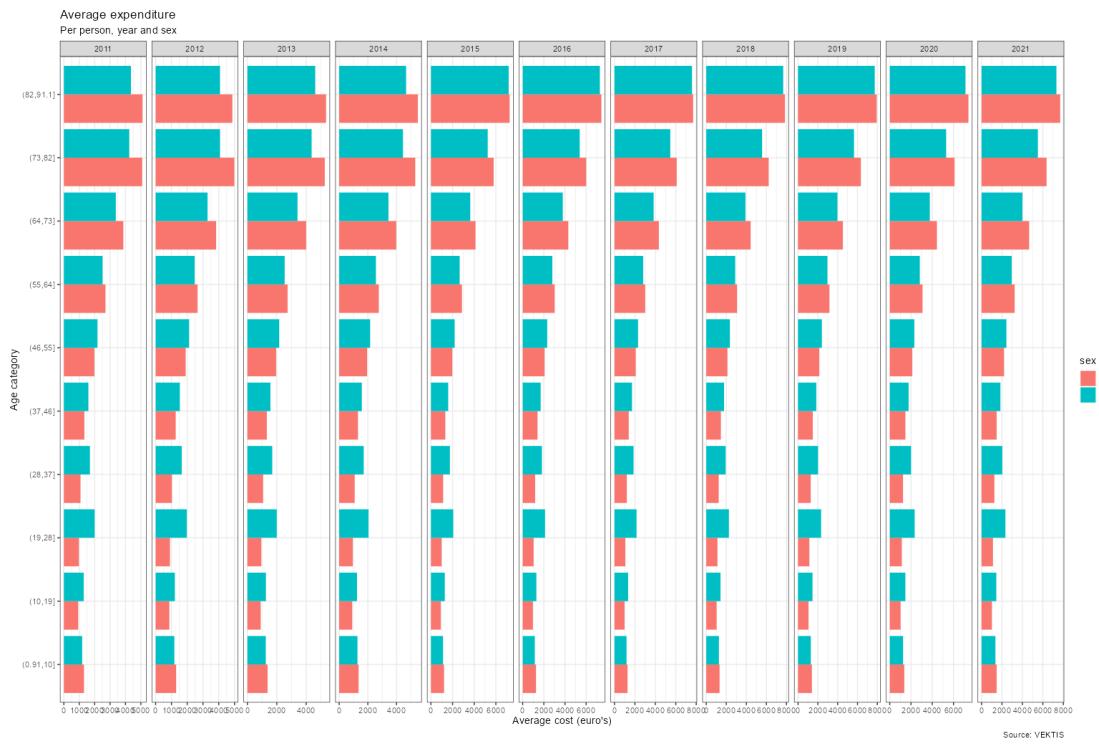


Figure 79. Average expenditure per person, year, and sex.

We already mentioned that ecological data only really lends itself to hunting trend reversals. So, the next graphs aim to show a trend reversal by plotting average healthcare expenditure for each year in which the last three years (2019, 2020 and 2021) are highlighted (Figure 80).

We split the data, again, by sex and age and here sees the same information as shown before: healthcare expenditure is a function of age in which the most expenses are made at age zero at ages above fifty. For women, there is a bump around the age of 30 which might be due to screening expenses, but this is not clearly shown in the data. Also, 2019-2021 have been amongst the costliest years, but 2019 was higher than 2020 or 2021, and especially for older age. It seems that at high age, the years start to diverge more in terms of average cost and that 2020 and 2021 were less expensive than 2019 (Figure 81).

Figures 82 - 84 show that specialist medical care is by far the most expensive category, and that 2020 was less than 2021 or 2019. The year 2021 showed the highest expensive in the past ten years for specialist medical care.

Last, but not least, we checked for the relationship between the number of insurance years and the cost for specialist medical care, considering year, sex, and the number of Dutch citizens (Figure 85). The expenditure peak is of course a function of the number of years that somebody is insured but did not show a clear trend reversal for 2020 or 2021.

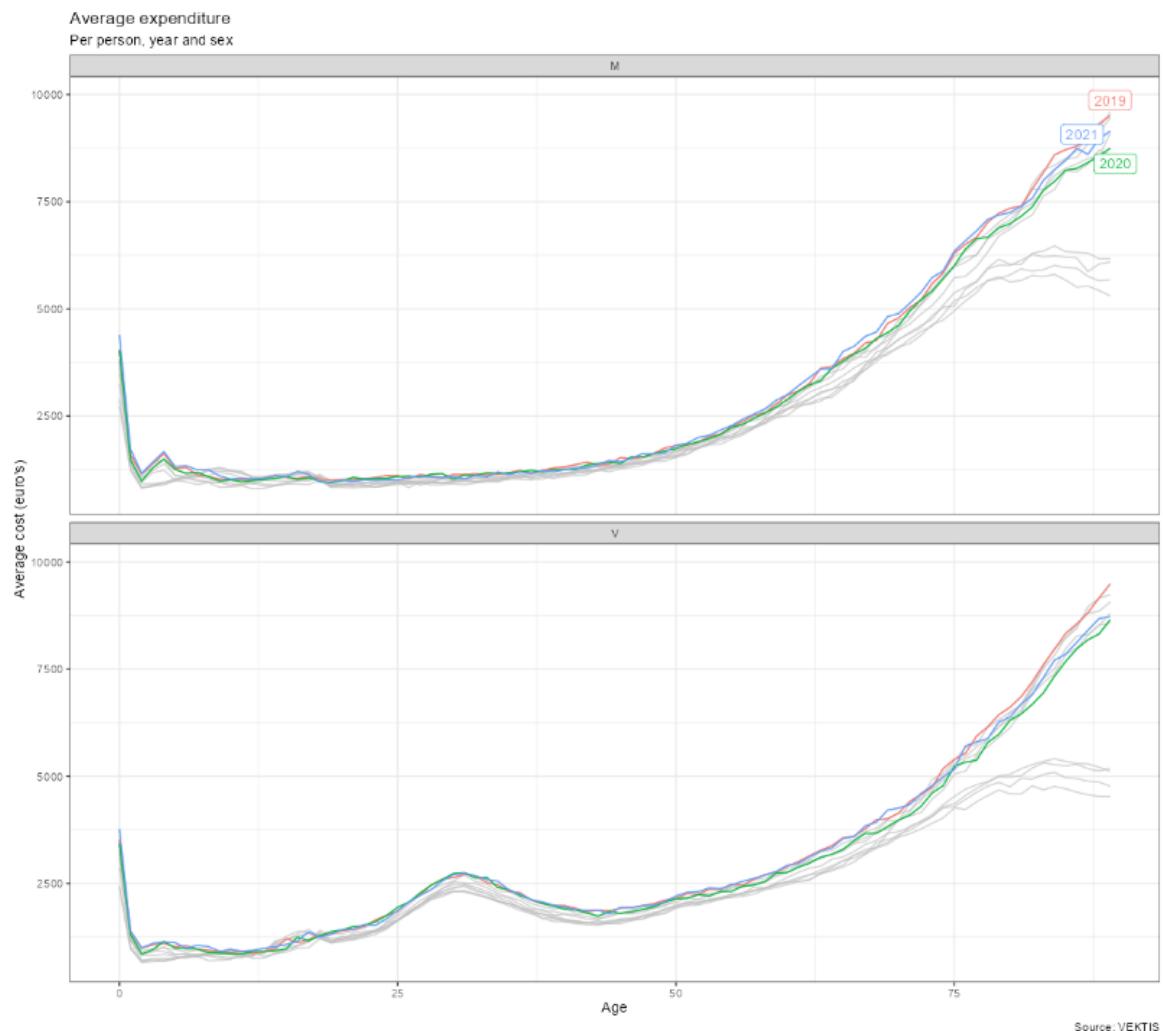


Figure 80. Average expenditure per person, showcasing per age, year, and sex highlighting 2019, 2020 and 2021.

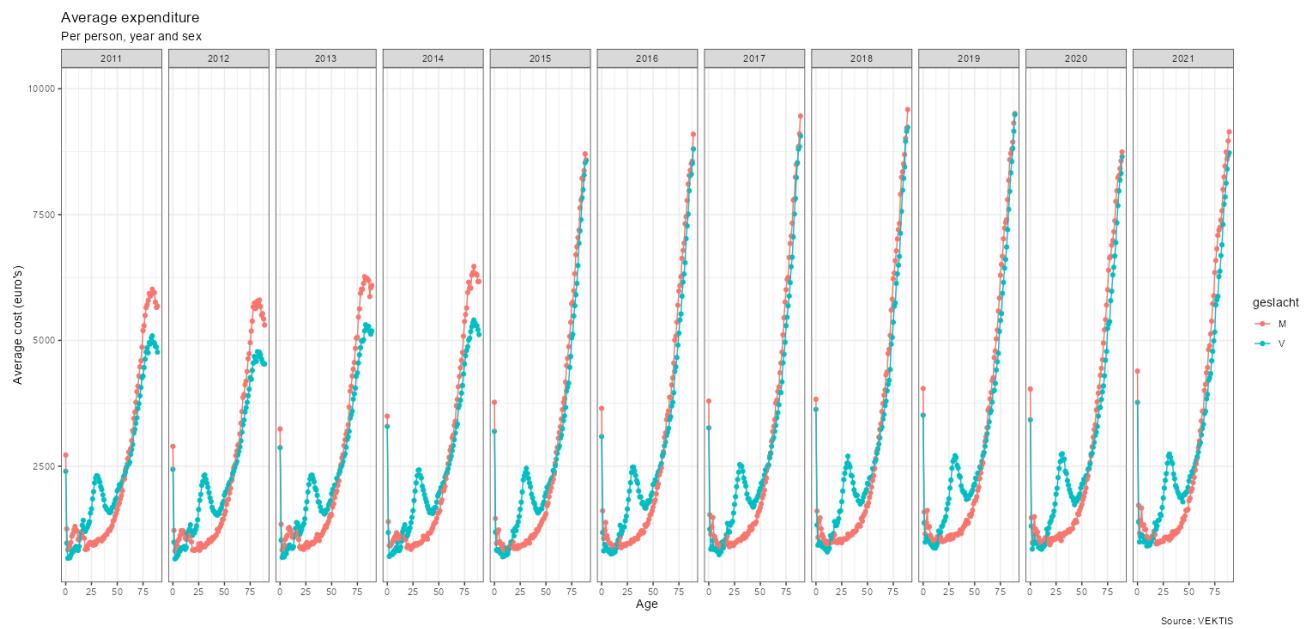


Figure 81. Average expenditure per person, showcasing per age, year, and sex highlighting 2019, 2020 and 2021.

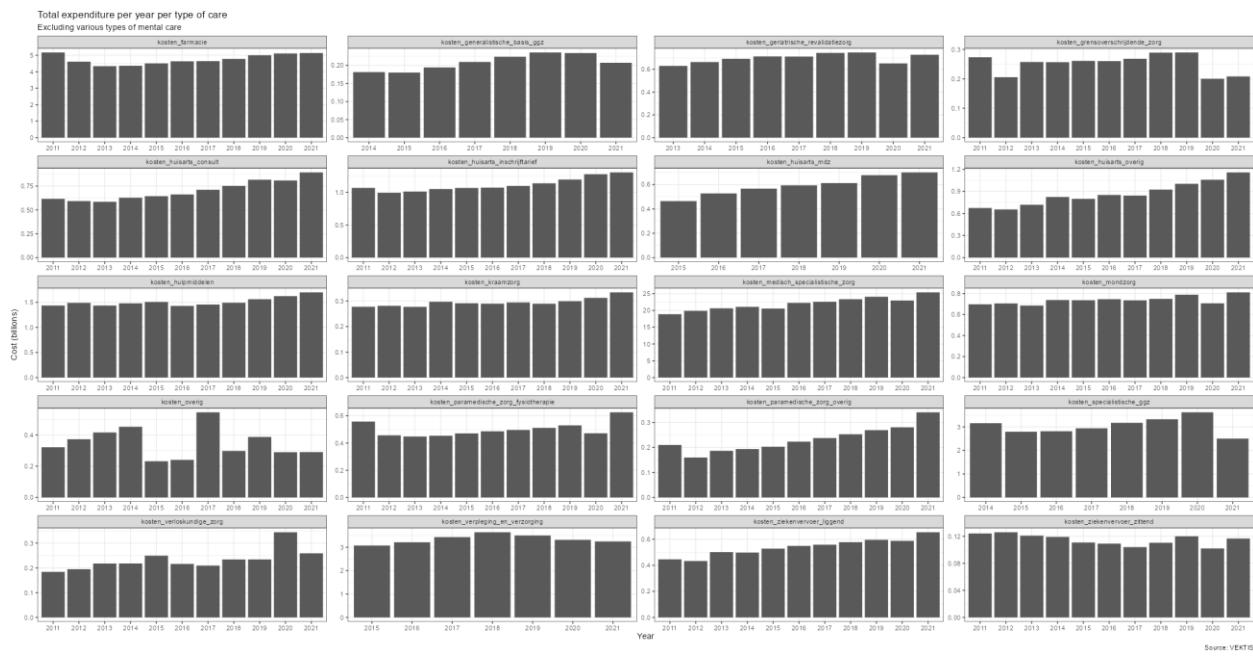


Figure 82. Total expenditure per year and per type of care.

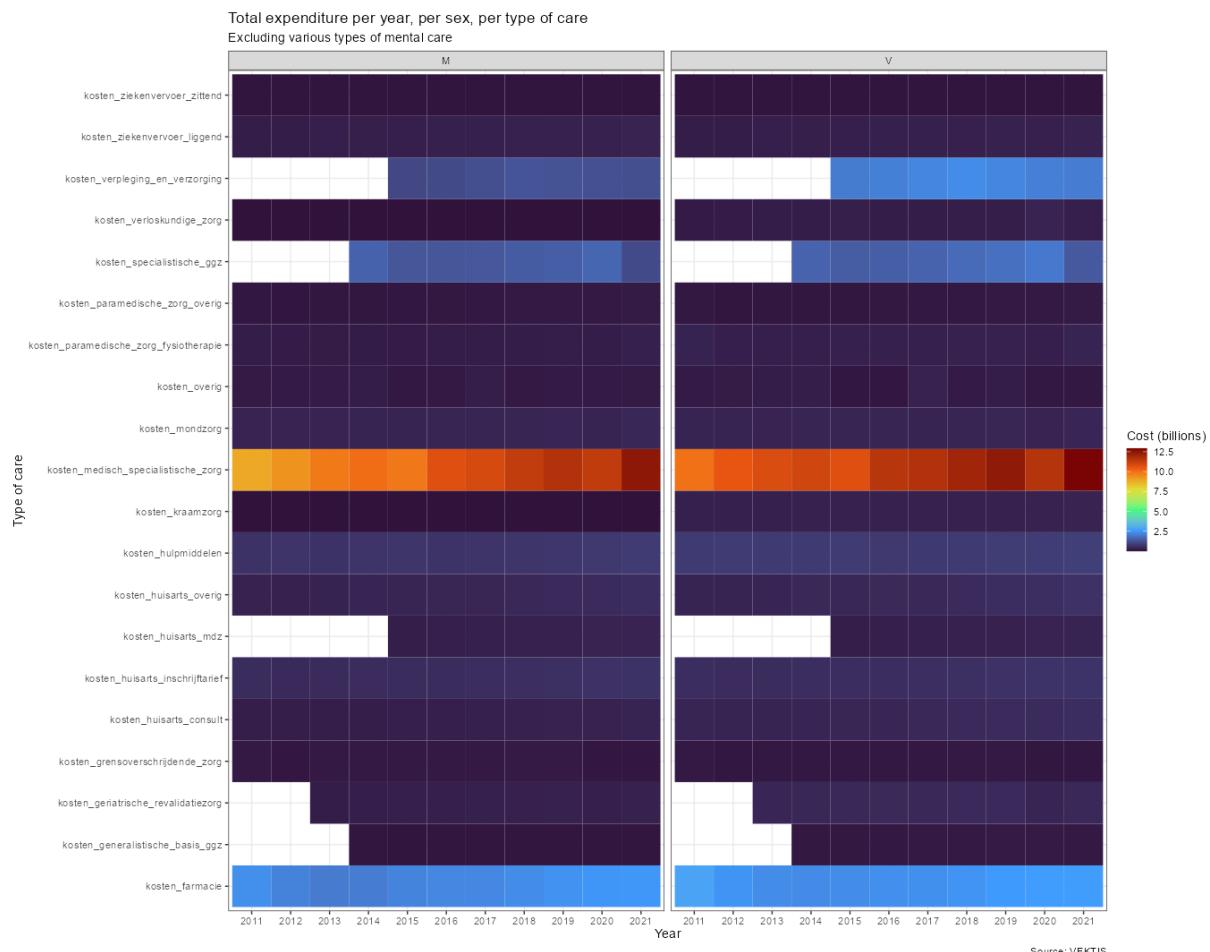


Figure 83. Total expenditure per year, per sex, per type of care (excluding various types of mental care).

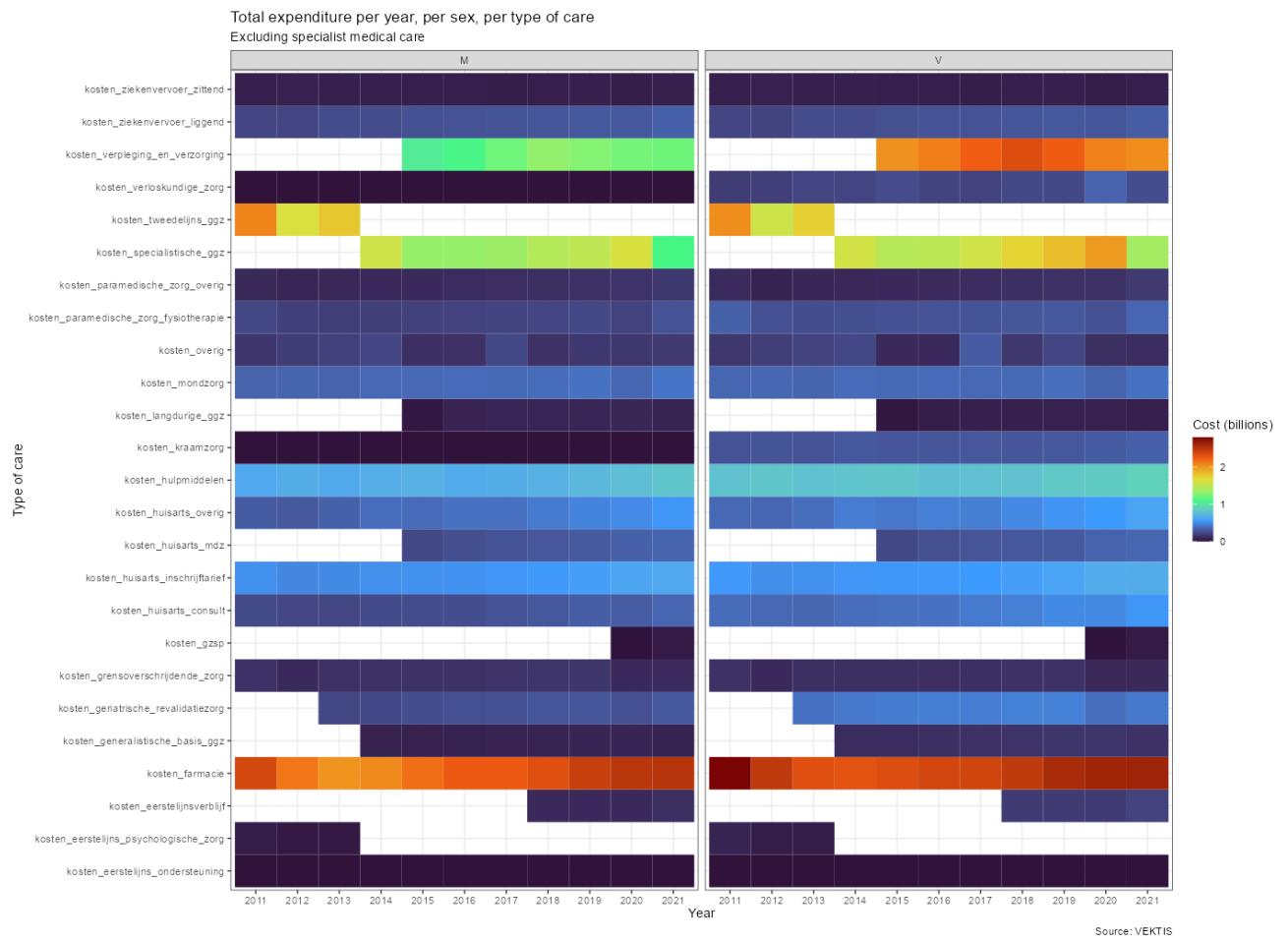


Figure 84. Total expenditure per year, per sex, and per type of care (excluding specialist medical care).

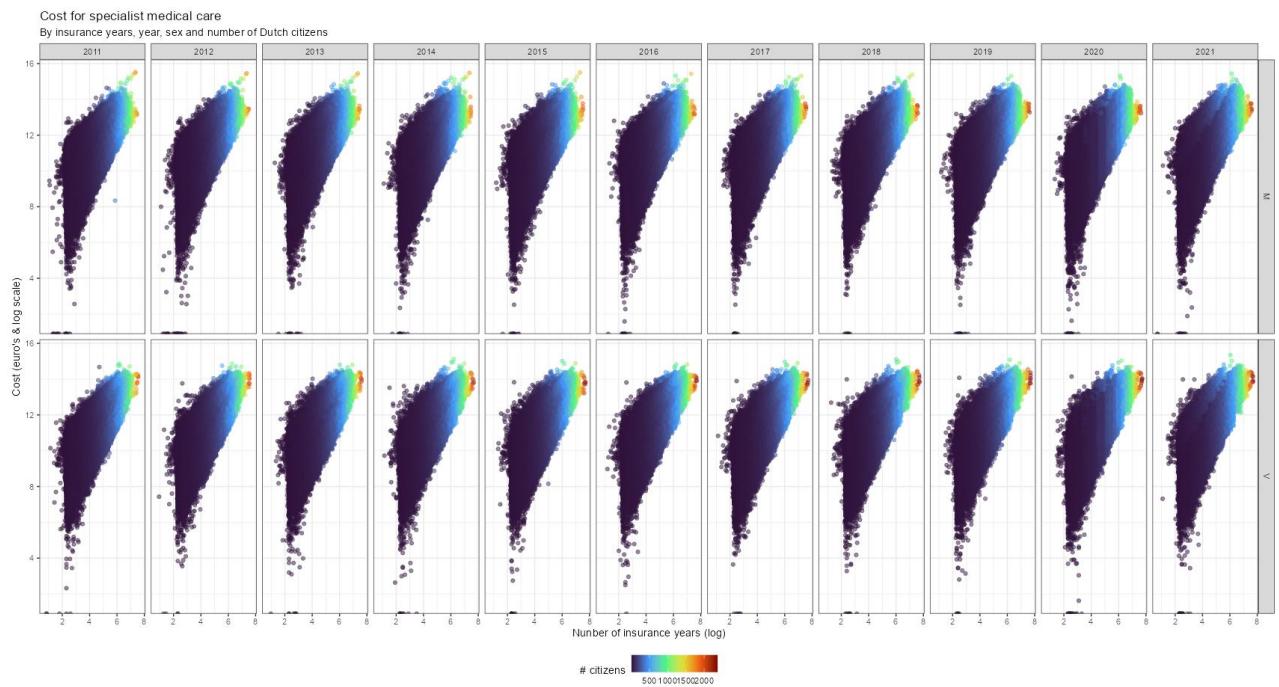


Figure 85. Cost for special medical care by insurance years, year, sex, and number of Dutch citizens.

We already mention that the OpenDIS data has more clinically informative data, but not on a postal level. For this project, the OpenDIS data was especially envisioned to detect metastatic cancer patients as a start for a more sophisticated analysis. However, this would actually be less sensitive data than the IKNL data we had envisioned. Hence, we will use the OpenDIS data from an ecological view to look for trend reversals.

To start, we want to look at the number of patients per healthcare product and diagnostic description, using the search term: “tumoren” (tumors). Easy to see is how benign tumors and mamma tumors (tumors of the breast) stand out (Figure 86).

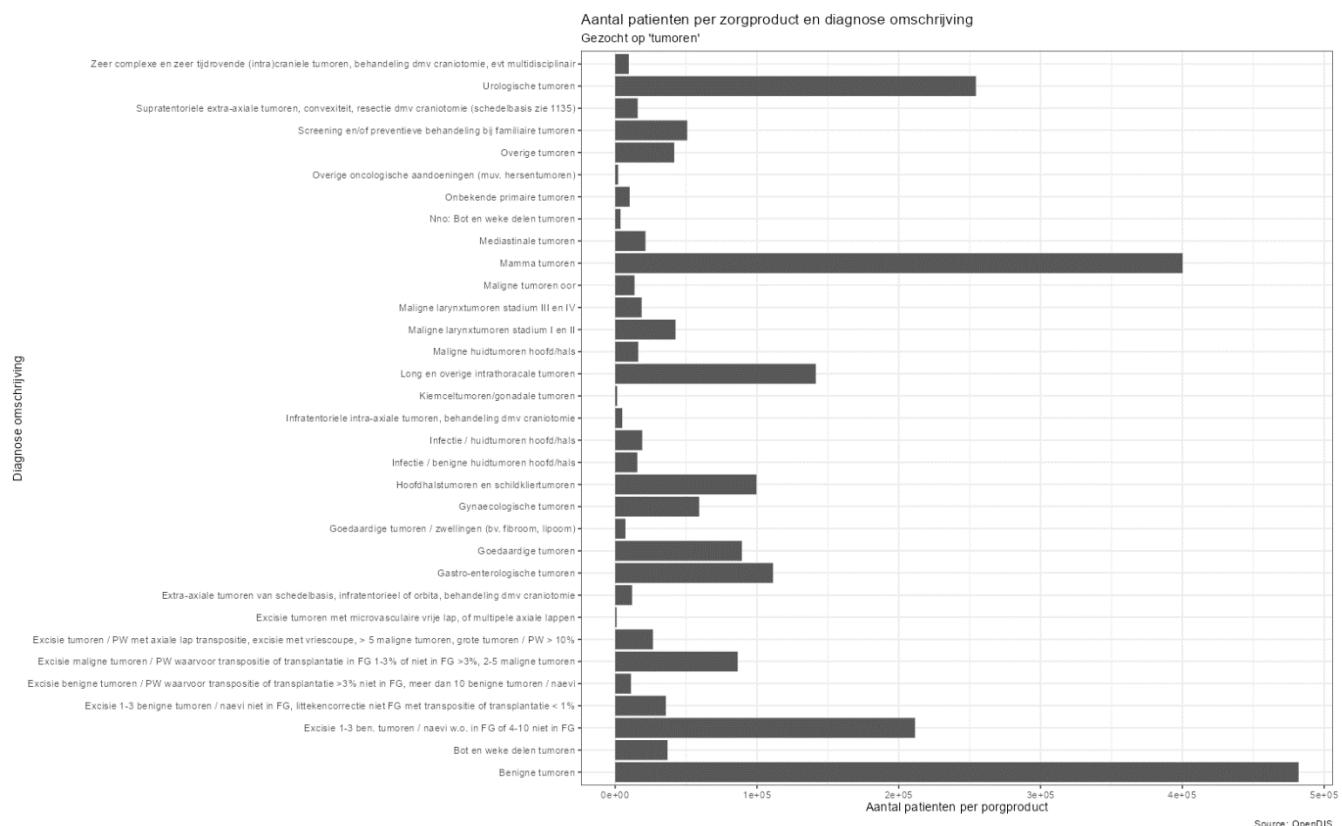


Figure 86. Number of patients per healthcare product and diagnosis (used ‘tumoren’ as search term).

Next, we want to see how much information OpenDIS has on screening which is shown in Figure 87. Clear to see is that OpenDIS does not have data on breast cancer or cervical cancer screening, but only some data on colon cancer. Looking at the specific healthcare products that are related to colon cancer, we see a clear trend reversal for 2020 (Figure 88). This is not really in line with the IKNL data which showed no real trend reversal for colon cancer. Because this is ecological data, we are quite certain where the difference originates from. To compare the screening data for colon to another form of screening, we also looked at screening for heritable forms of cancer for which breast cancer may indeed be a part of (Figure 89). Also here, we see a trend reversal but is not large.

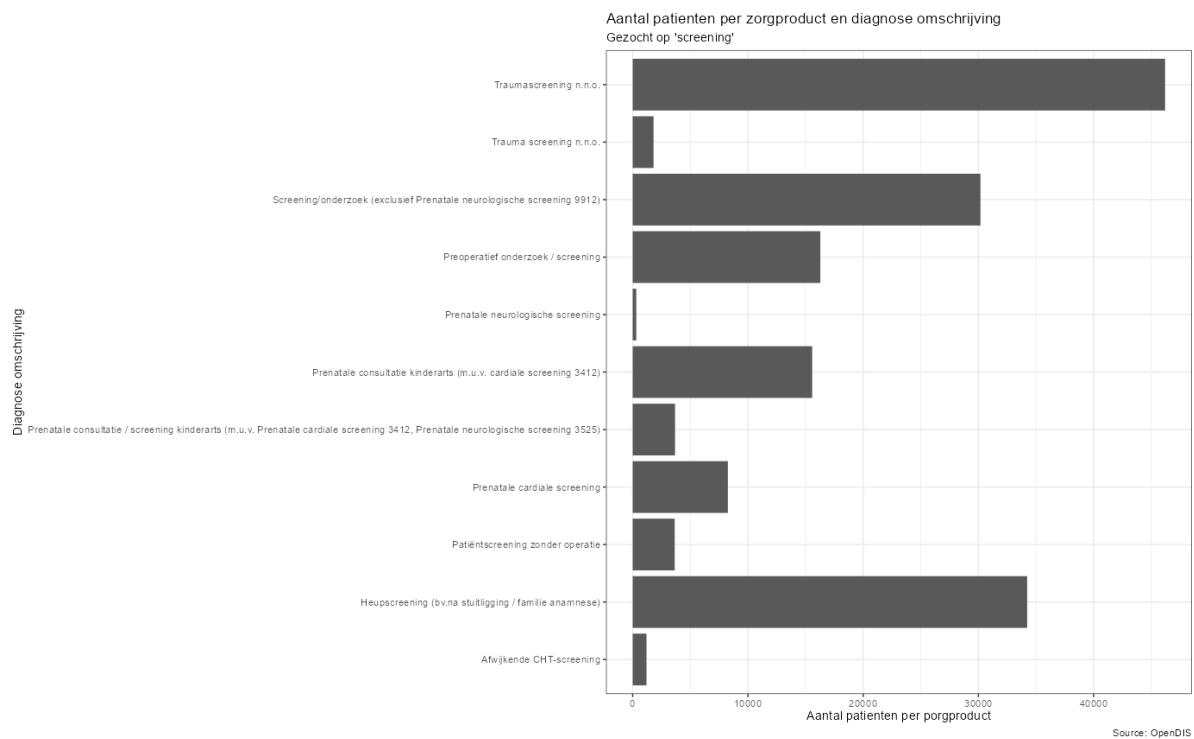


Figure 87. Number of patients per healthcare product and diagnosis (used 'screening' as search term).

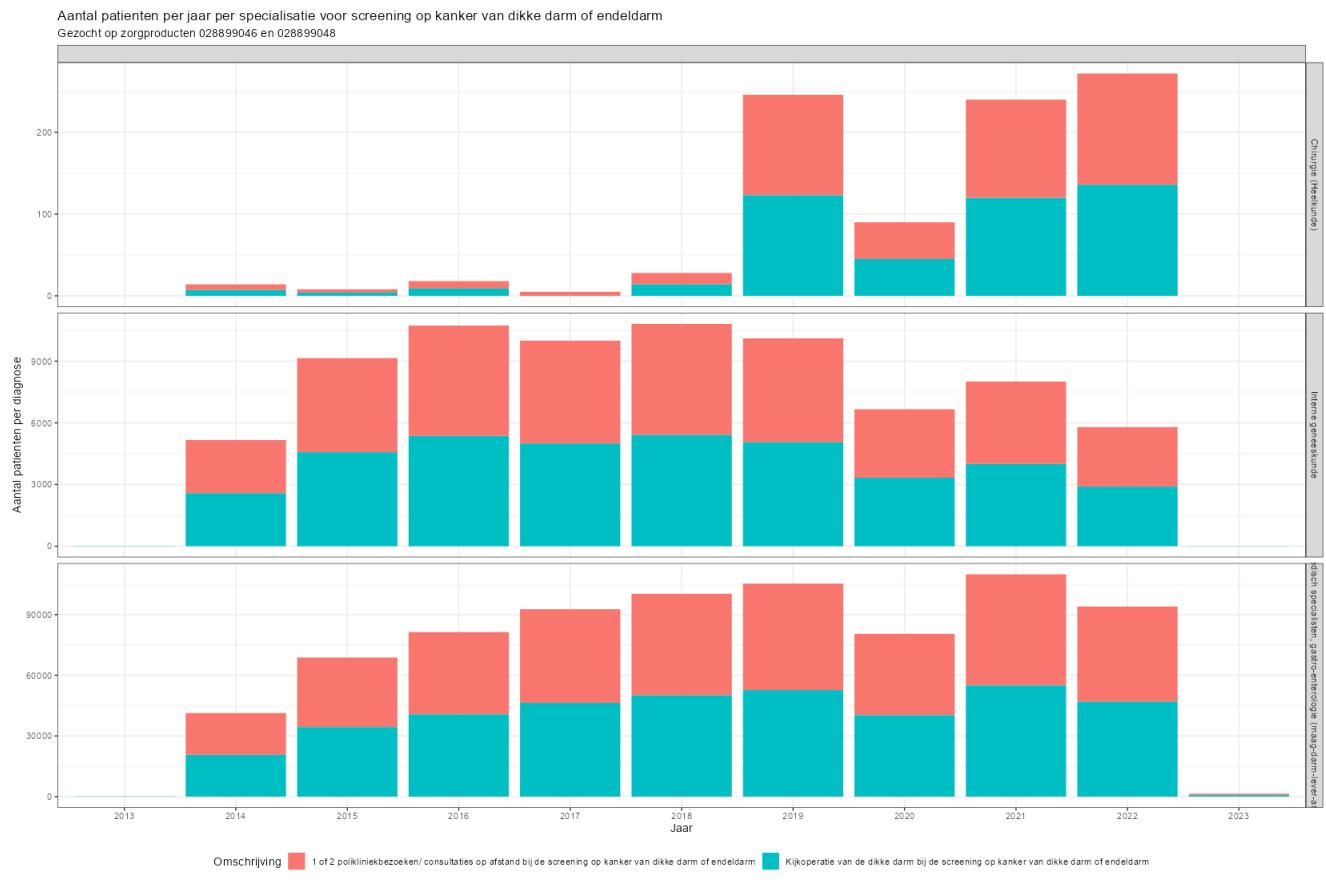
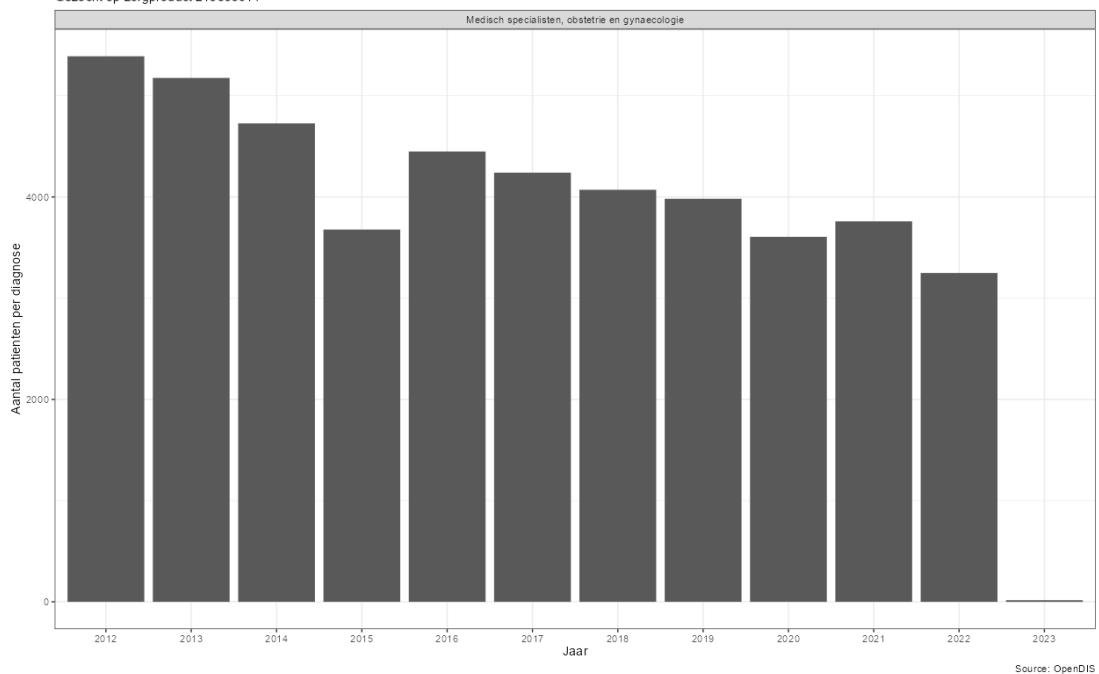


Figure 88. Number of patients for colon cancer screening per year per healthcare specialization.

Aantal patienten per jaar per specialisatie voor screening op erfelijke vormen van kanker
Gezocht op zorgproduct 219899014



Source: OpenDIS

Figure 89. Number of patients per specialization for screening on heritable cancer.

An additional search for ‘research’ did not provide any additional information but shows that 2020 was a trend reversal for colorectal cancer screening (Figure 90). The same goes for a search using the terms ‘bevolkingsonderzoek’ and ‘screening’ (Figure 91).

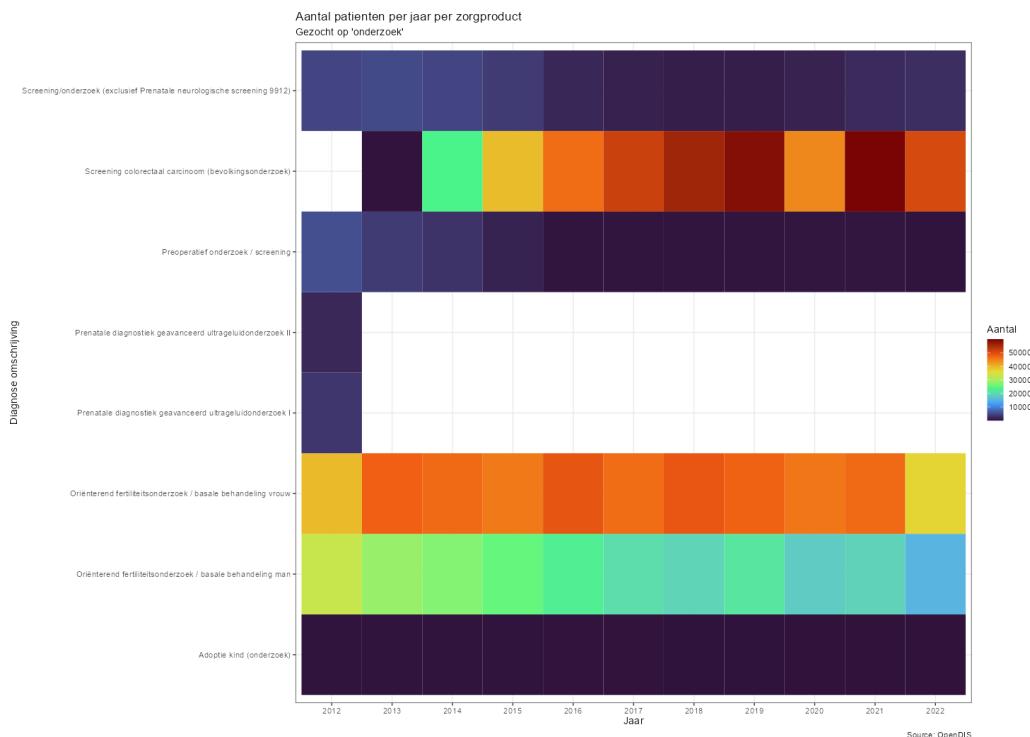


Figure 90. Number of patients per year per healthcare product, using the search term ‘onderzoek’.

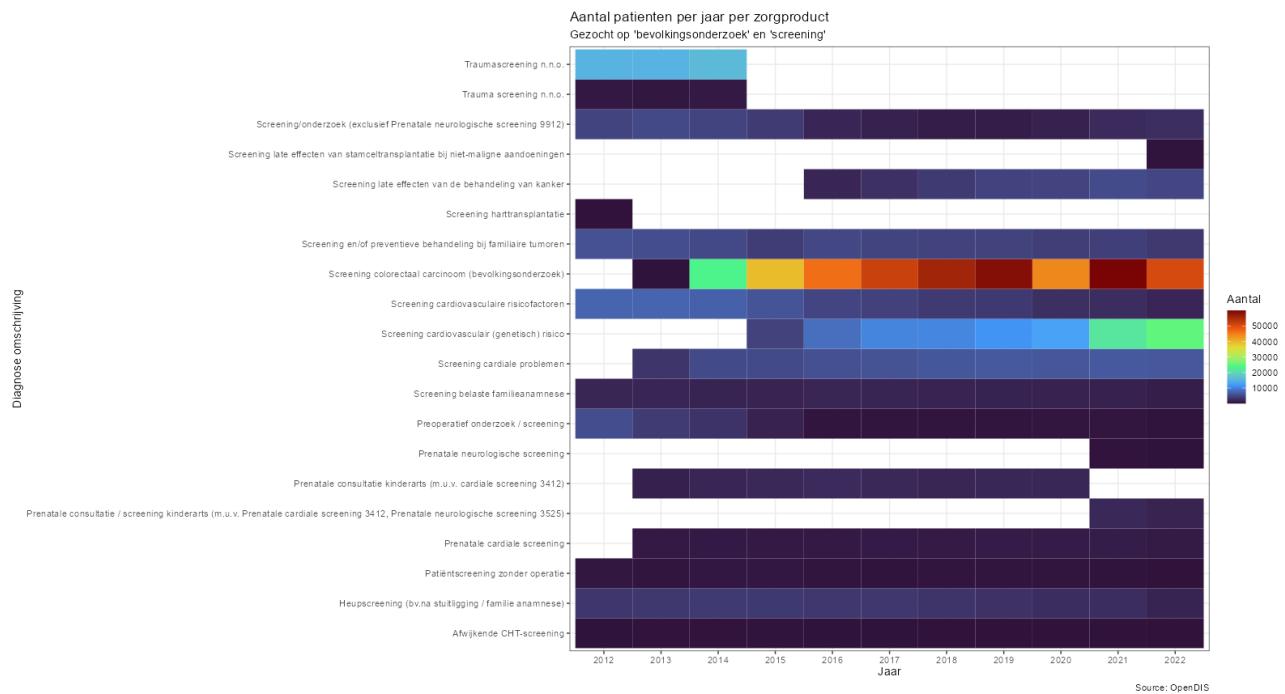


Figure 91. Number of patients per year per healthcare product, using the search terms 'bevolkingsonderzoek' and 'screening'.

Next, we looked more closely at the OpenDIS data for cancer by searching for any healthcare product that contained the words 'kanker', 'carcinoom' or 'tumor' in their description (Figure 92). Using a heatmap, we can see that there are many tumors recorded in OpenDIS but is not clear to see that for each of them 2020 showed a trend reversal.

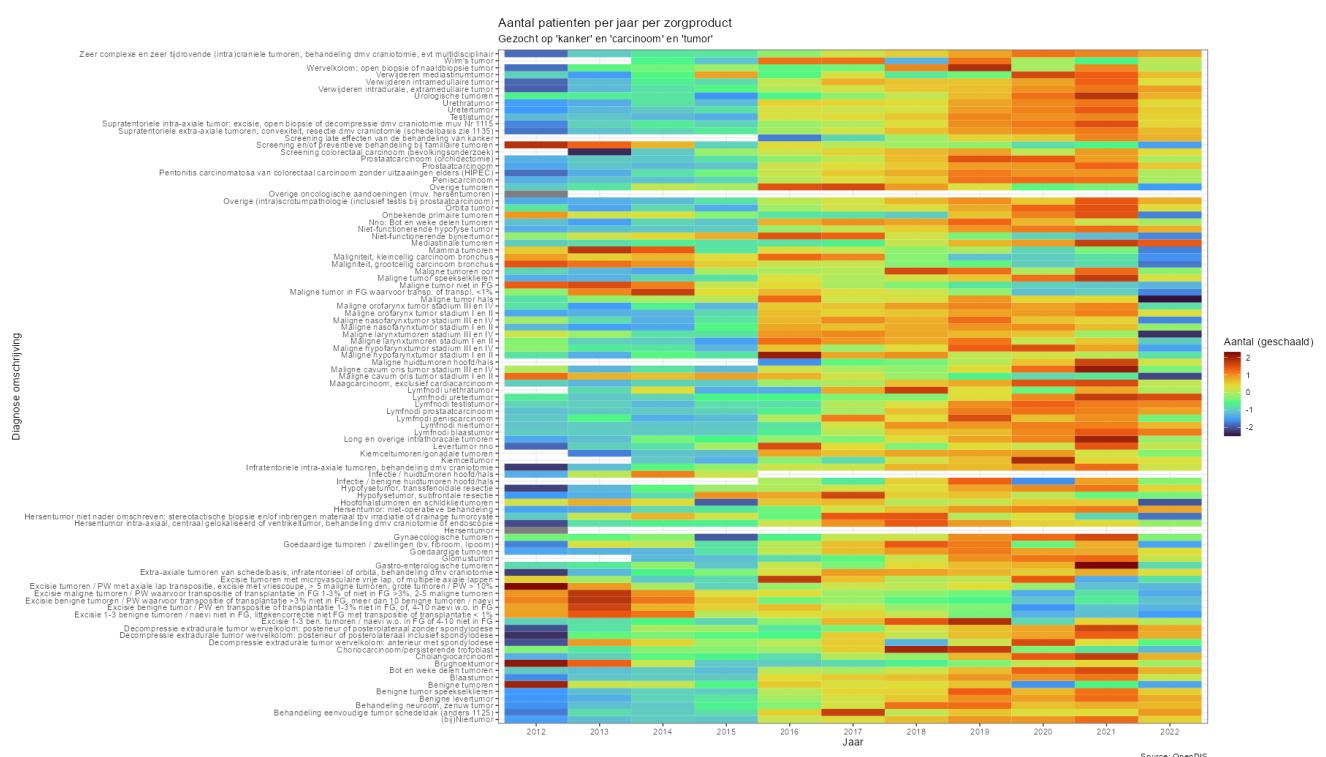


Figure 92. Number of patients per healthcare product, using search terms 'cancer' and 'carcinoom' and 'tumor'.

If we look deeper by using the word ‘tumoren’ we can see trend breached for several tumors based on standardized values (Figure 93).

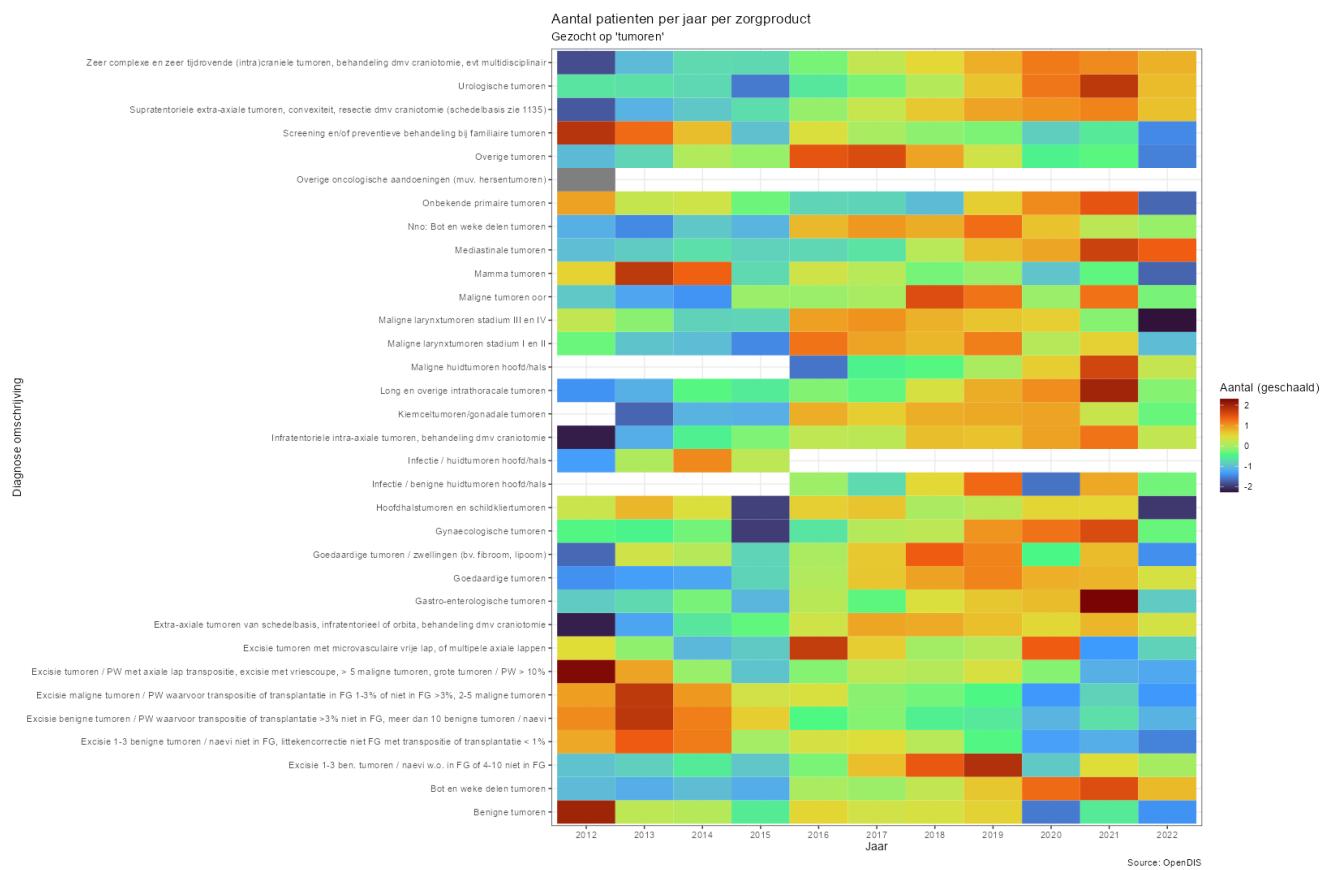


Figure 93. Number of patients (scaled) per year per healthcare product, using the search term ‘tumoren’.

The OpenDIS dataset is a large dataset containing all the DBC information, split by type of diagnosis and care received, and who provided that care. Hence, we will split any further analysis by type of cancer (breast, cervical and colon). First, we looked at healthcare products associated with breast cancer tumors (‘mamma tumoren’). Using standardized values once again, it does not directly become clear that 2020 is an overall trend reversal (Figure 94). Despite seeing changes from 2019 to 2021 in certain categories, we cannot say that these changes did not occur in the years before. If we dive deeper into the level of care received (Figure 95), we can see that there is a lot of radiotherapy involved, but that was already declining.

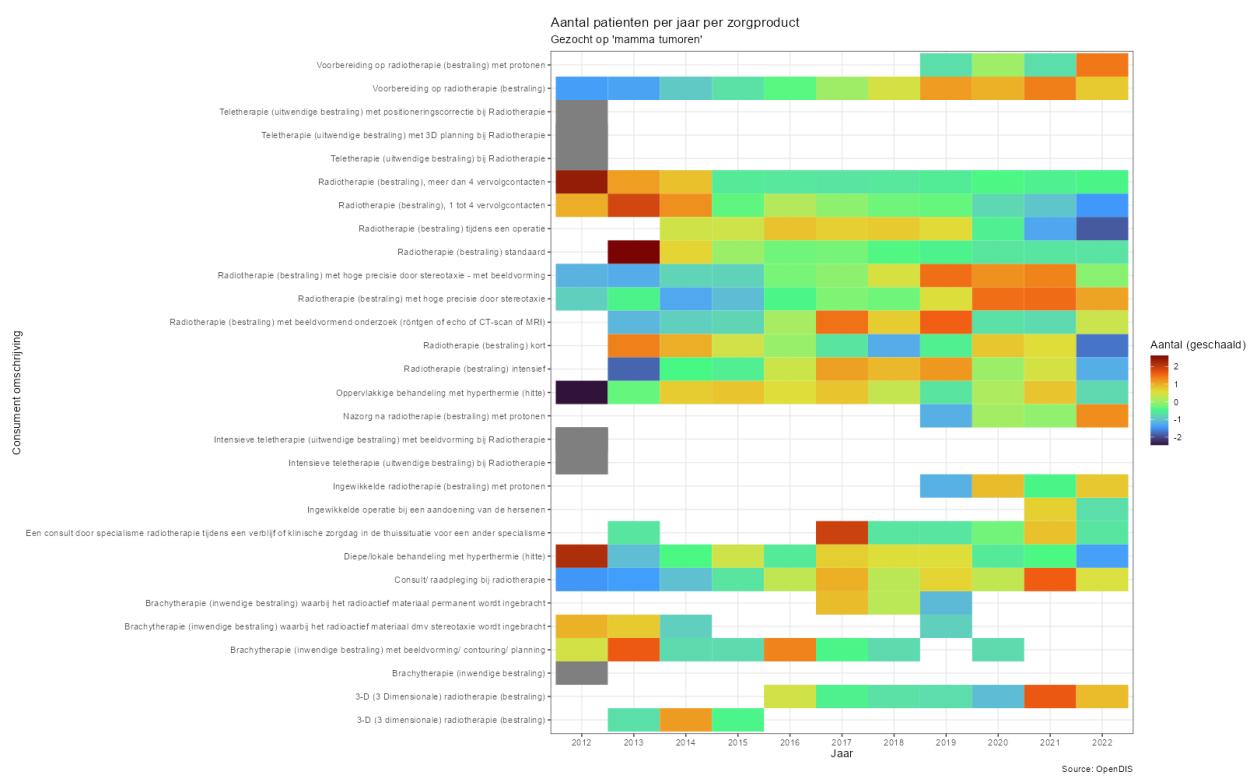


Figure 94. Number of patients (scaled) per year per healthcare product, using the search term 'mamma tumoren'.



Figure 95. Number of patients diagnosed with breast cancer per type of treatment.

Perhaps our analysis makes more sense if look at all the types of care that are received for patients diagnosed with breast cancer and rank them across the years. Here, the highest rank is specified by the number one and the lowest rank equals the number twelve. Hence, the higher the rank, the higher the number of patients who received that product of care. Looking at Figure 96 it is very difficult to see a clear trend reversal. Once again, the ranking per category does not seem to follow a slope across the years. We also do not see that 2020 is necessarily a year in which less treatment was given. For instance, the year 2023 is not complete and so it makes sense that for almost all categories the ranking is lowest, because there are months missing. The year 2022 also shows quite a low ranking.

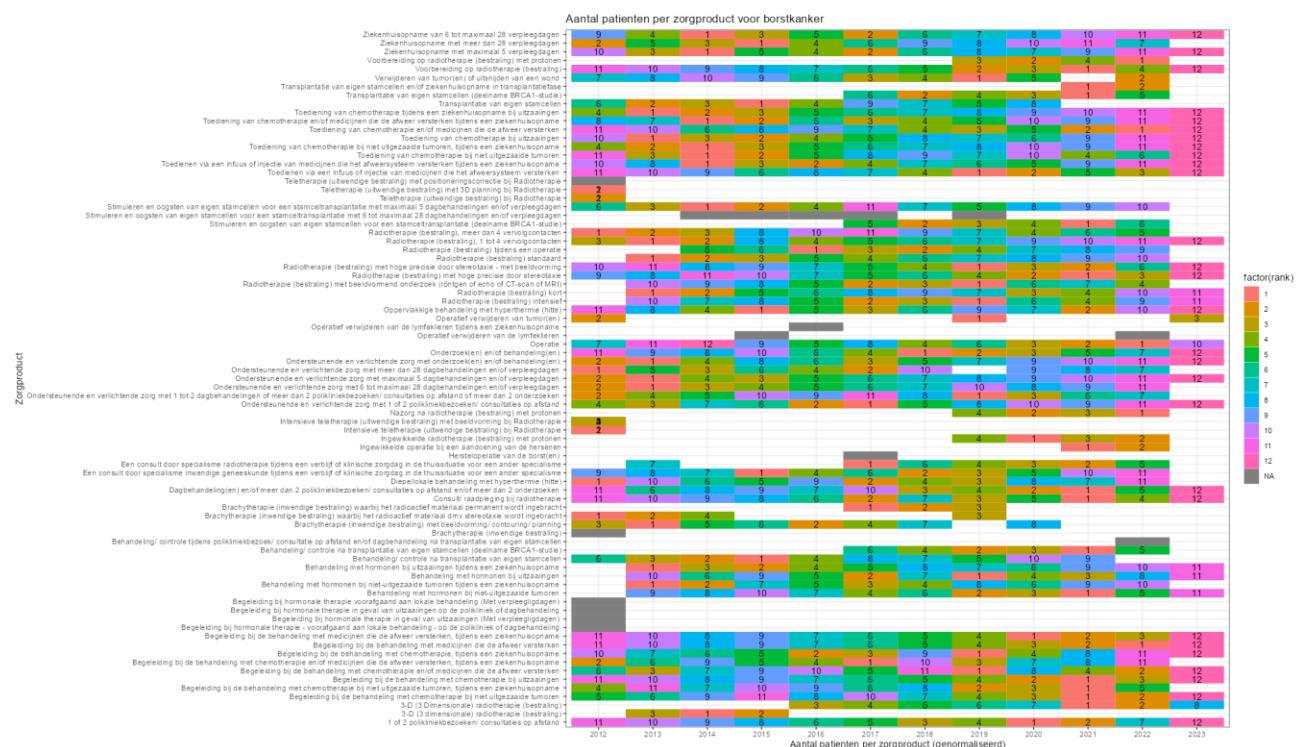


Figure 96. Number of patients (scaled) diagnosed with breast cancer per type of treatment. The colors show the rank: the higher the rank number the more patients included in that year per type of treatment.

From the OpenDIS data it is not immediately clear that 2020 showed a severe change in treatment provided for breast cancer patients. Next, we could focus on changes within the care provided. Doing so yields Figure 97 which shows no trend reversal in the number of DBCs recorded for breast cancer treated via surgery ('heelkunde') or internal medicine ('interne geneeskunde'). Looking more specifically at certain codes for healthcare products related to metastatic breast cancer does not show any trend reversal (Figure 98).

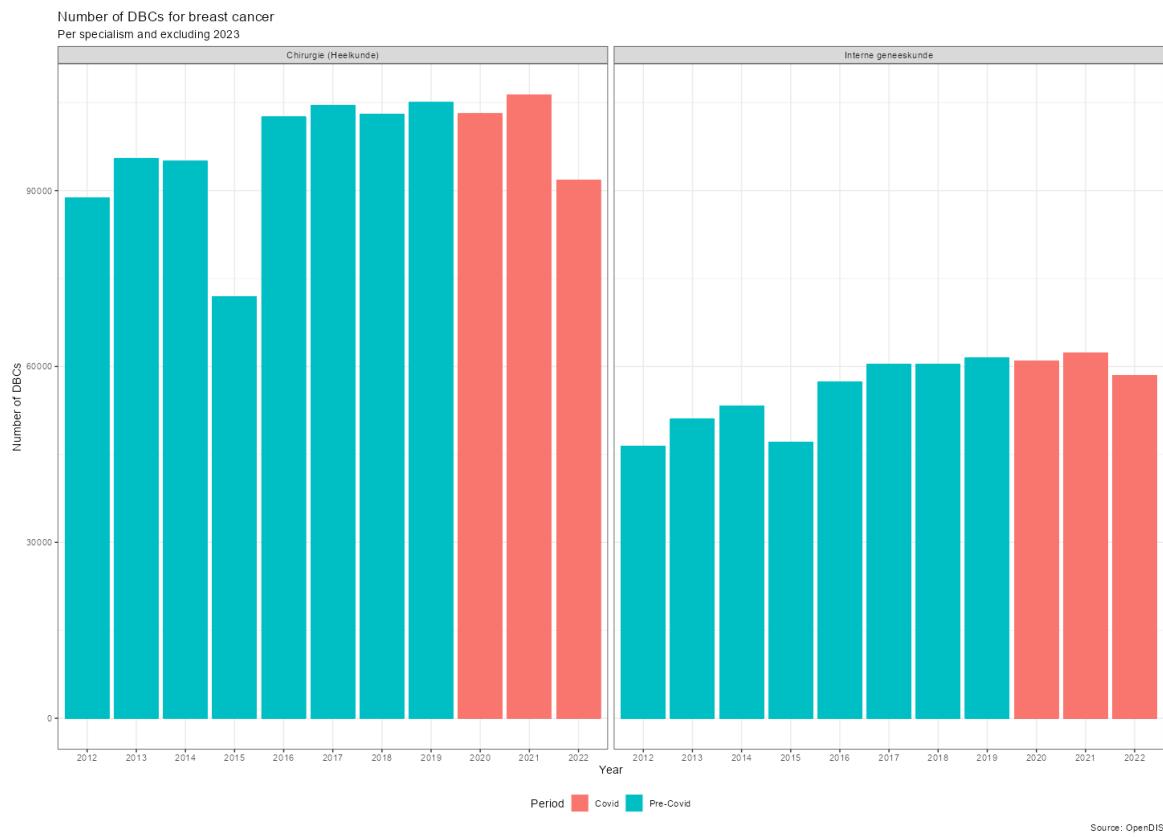


Figure 97. Number of DBCs for breast cancer, per specialism and period.

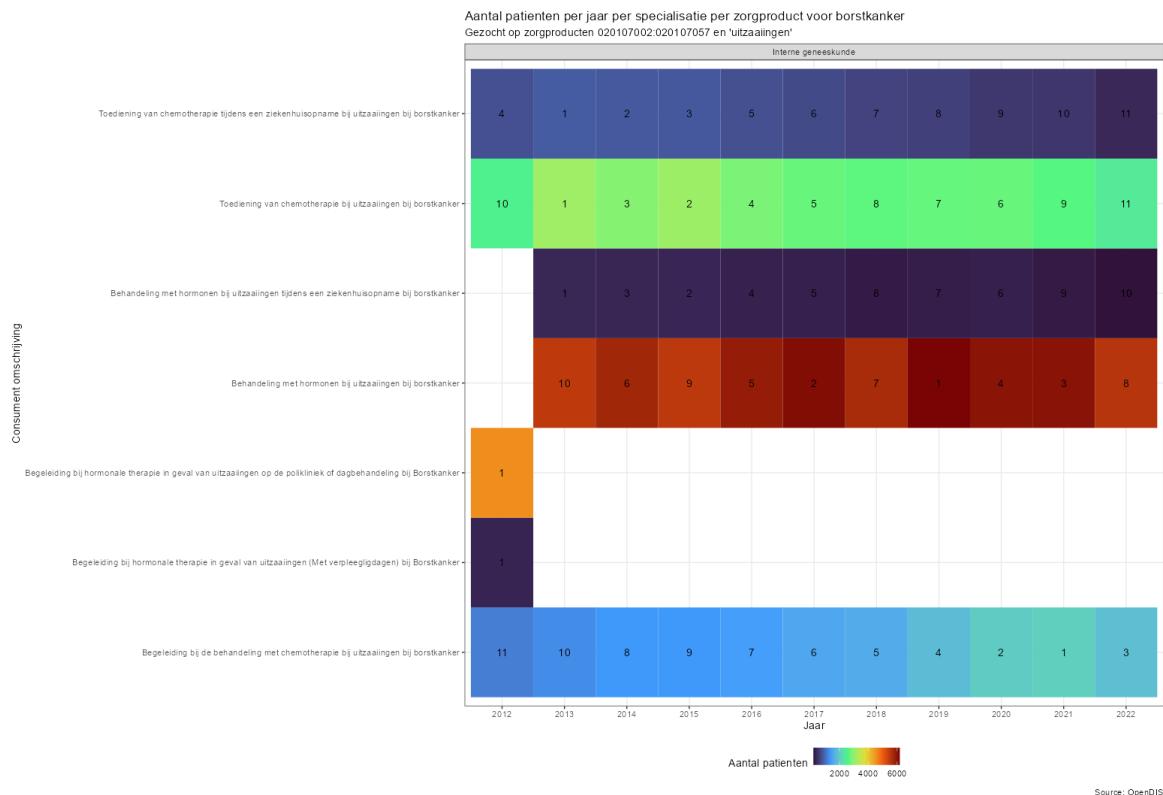


Figure 98. Number of patients per year per specialist per healthcare product for breast cancer using healthcare products '020107002:020107057 and the search term 'uitzaaiingen'.

Next, we will focus on colon cancer patients by searching for care products that contain the words ‘darmkanker’ and ‘kanker van dikke darm’ in their description. Looking at Figure 97 we see once again the dip in colorectal screening but do not see any trend reversal in the number of colon cancer diagnoses. However, when looking at the number of DBCs for colon cancer specified by type of care, we do see a 2020 dip by medical specialist care (Figure 100).

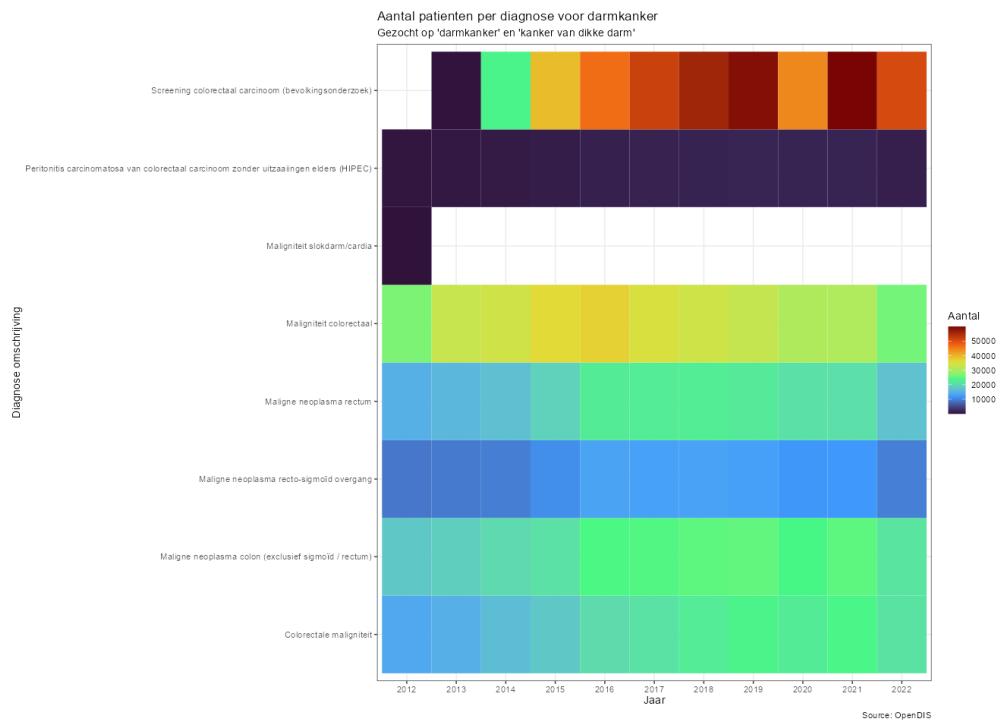


Figure 99. Number of patients per year per diagnosis for colon cancer, using search terms for ‘darmkanker’ and ‘kanker van dikke darm’.

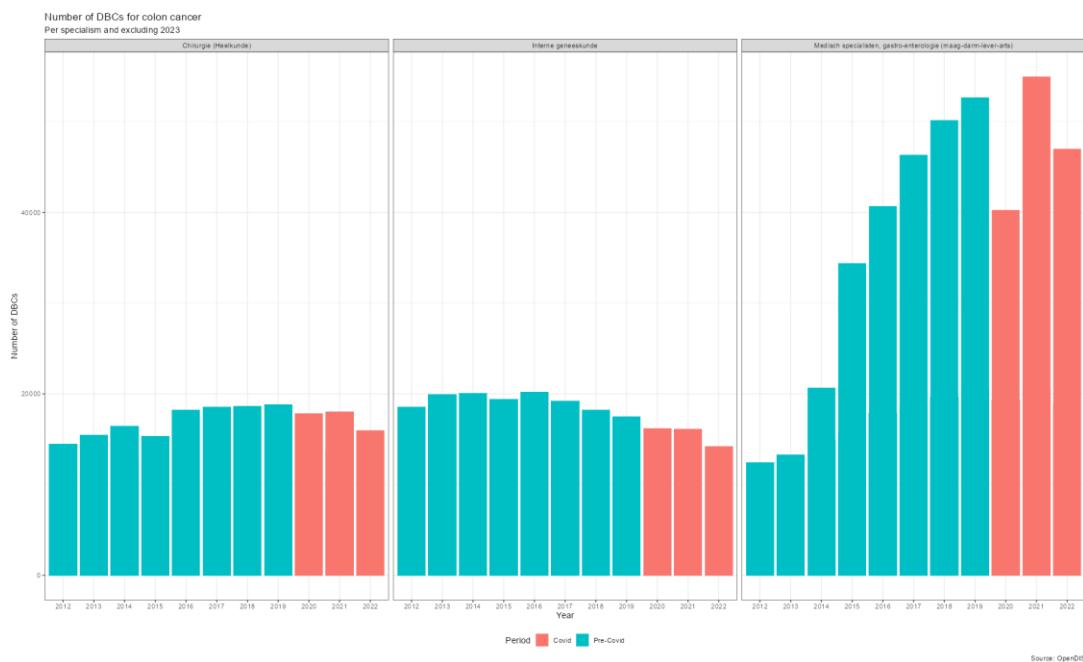


Figure 100. Number of DBCs for colon cancer per specialist and per year highlighting the pre-Covid-19 and Covid-19 period.

There are many different types of colon cancer and searching for ‘dikkedarm-endeldarmkanker’, ‘darmkanker’ or ‘kanker van dikke darm’ shows multiple different diagnoses. Plotting them per year and per type of care received does perhaps help in finding different trend reversals (Figure 101). Indeed, for certain cancer types the number of diagnoses dropped after 2020 to a level like 2015. It is unclear if the years between 2015 and 2020 were different in themselves to account for such a change.



Figure 101. Number of patients per diagnosis for colon cancer, per year, per specialist and per specific type of colon cancer.

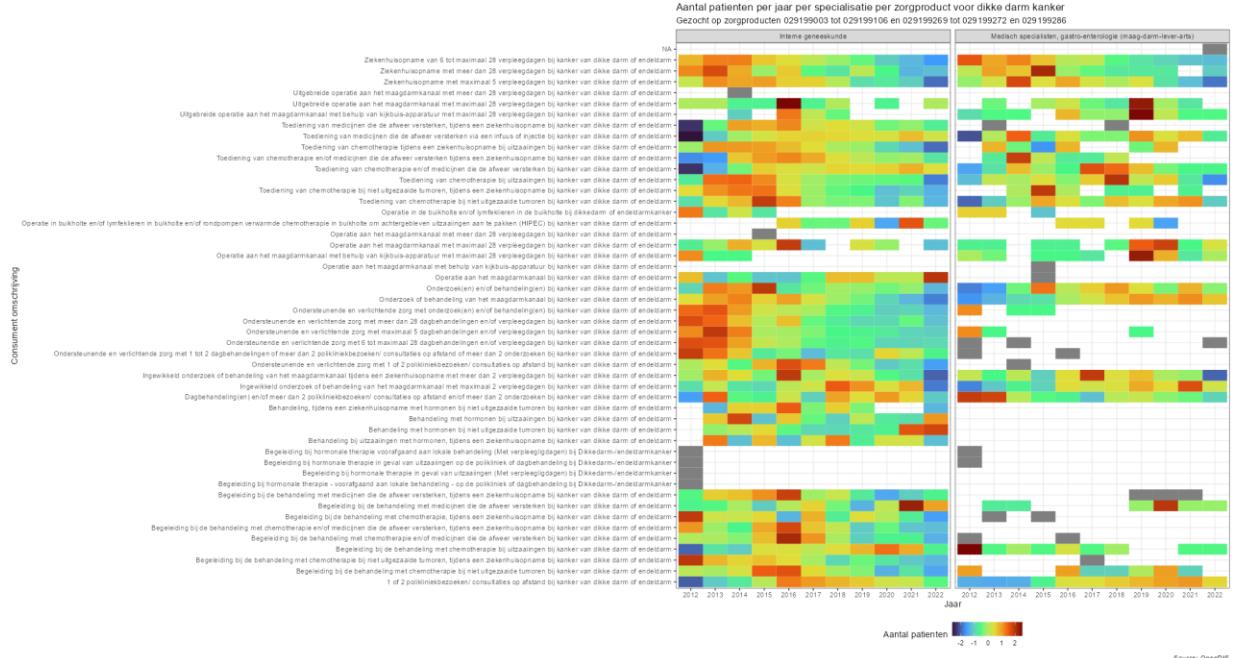


Figure 102. Number of patients per year, per specialist, per healthcare product for colon cancer, looking for specific healthcare products that have colon cancer in their description.

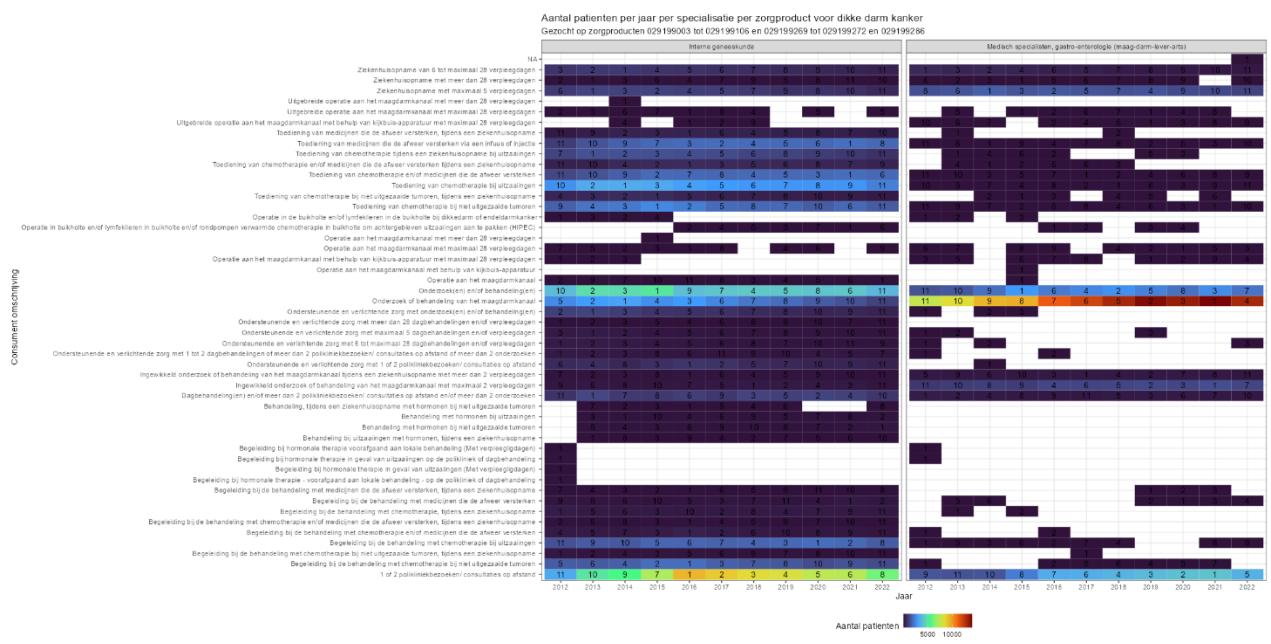


Figure 103. Number of patients per year, per specialist, per healthcare product for colon cancer, looking for specific healthcare products that have colon cancer in their description, ranked.

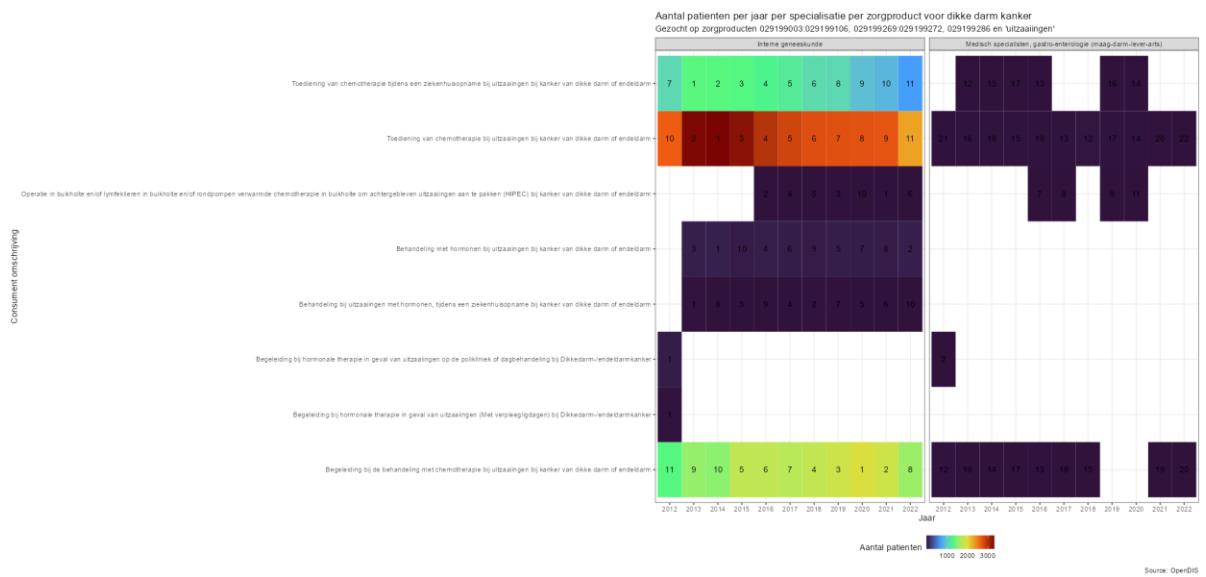


Figure 104. Number of patients per year, per specialist, per healthcare product for colon cancer, looking for specific healthcare products related to metastatic colon cancer, ranked.

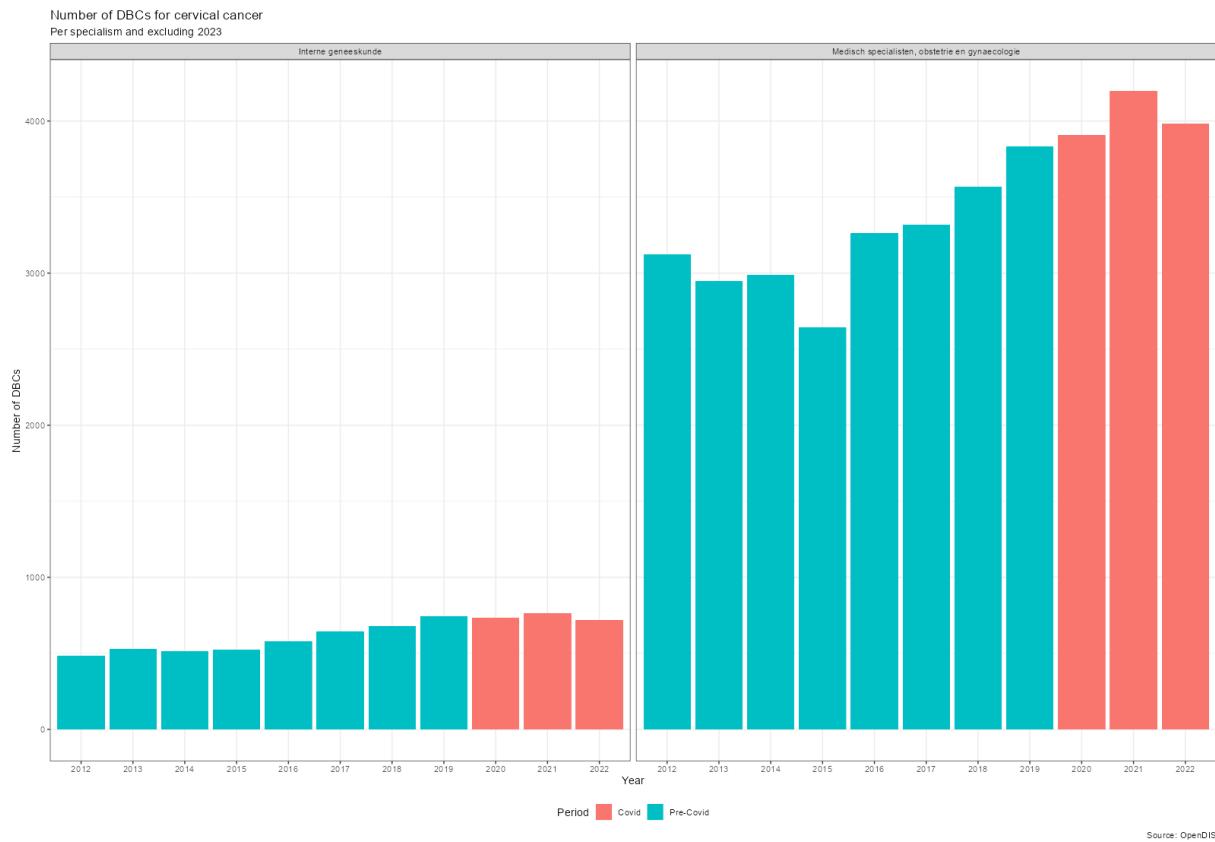


Figure 105. Number of DBCs per year, specialist, and Covid-period for cervical cancer.

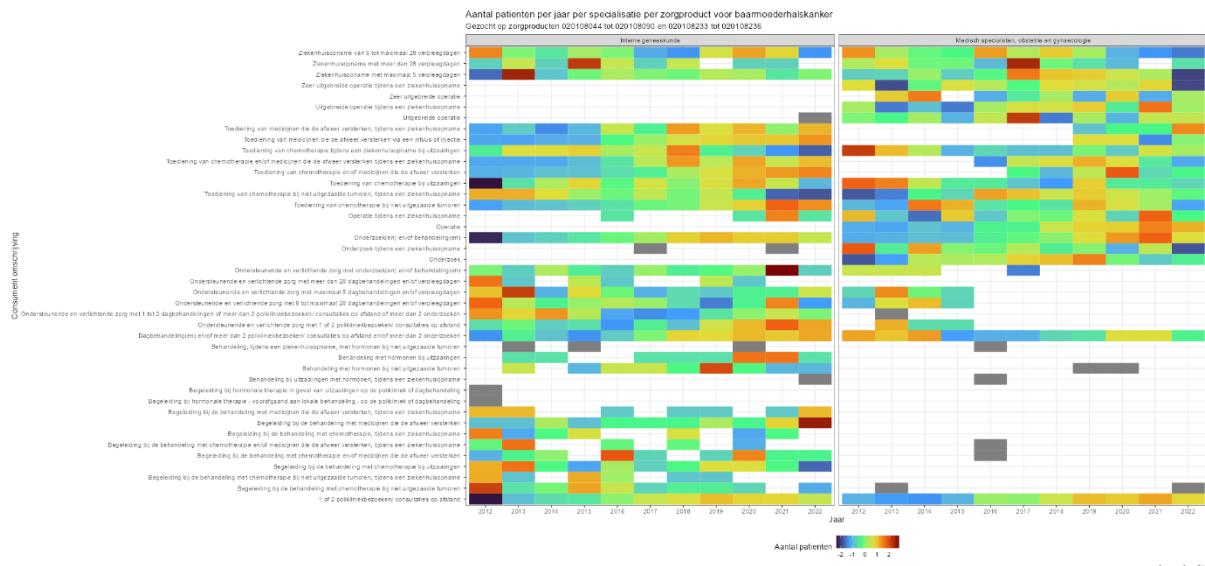


Figure 106. Number of patients per year, per specialist, per healthcare product for cervical cancer, looking for specific healthcare products that have cervical cancer in their description.



Figure 107. Number of patients per year, per specialist, per healthcare product for cervical cancer, looking for specific healthcare products that have cervical cancer in their description, ranked.

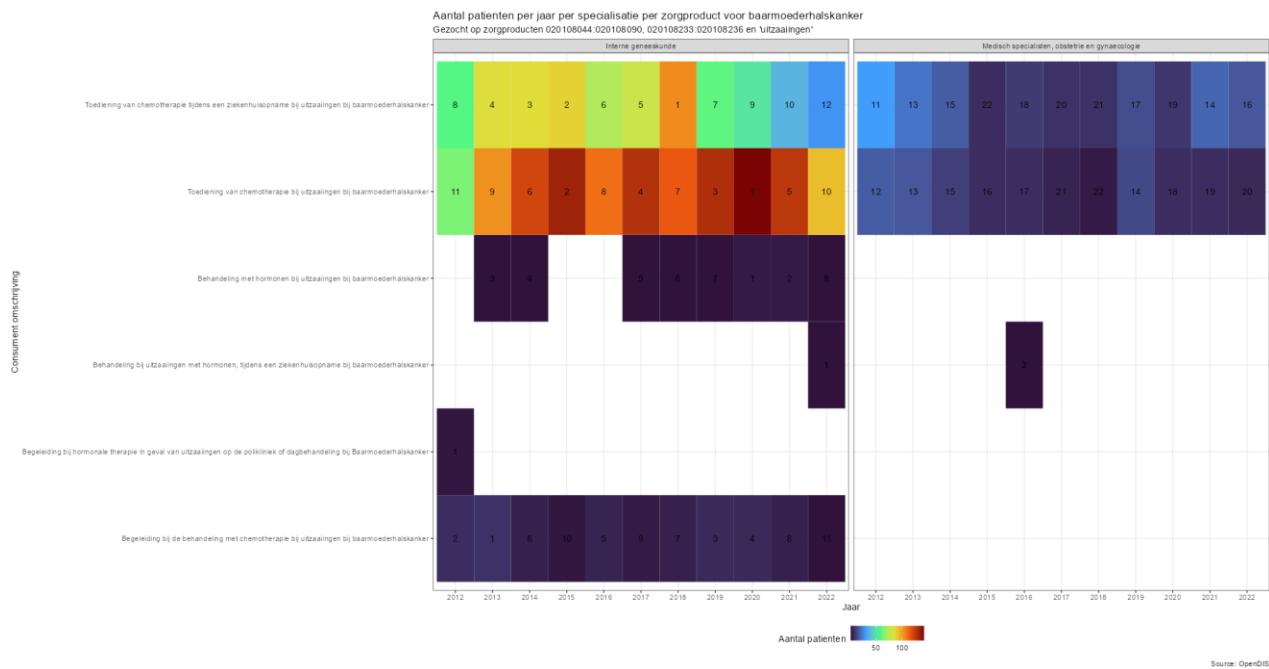


Figure 108. Number of patients per year, per specialist, per healthcare product for cervical cancer, looking for specific healthcare products related to metastatic cervical cancer, ranked.

Clinical prediction model for breast cancer

Rapid review of clinical prediction models for breast cancer.

From this rapid review, which was conducted in the week of September 4th to September 8th, [we collected 33 studies published between 1999 and 2023](#). Most of the studies describe the creation of a prediction model for breast cancer whilst some tried to validate known prediction models. Currently, the most widely used prediction models are PREDICT¹⁰⁷ and CancerMath.¹⁰⁸ The latter has several prediction tools available.

The key aim of this rapid review was to detect which factors would be most important when trying to predict survival, and if there are already prediction models available that we could use on predicting the average survival of the average group of patients diagnosed in the Netherlands via screening.

From the Netherlands, we only found two studies. One was a study describing the creation of a model estimating the annual risk of locoregional recurrence.¹⁰⁹ The other was a validation study for the CancerMath prediction tool.¹¹⁰ In this study it was shown that CancerMath predicts overall survival quite well in early breast cancer patients although the model is certainly not without its limitations. Nevertheless, the study gave us some confidence in using CancerMath.¹¹¹ In addition, we used PREDICT since it had quite

¹⁰⁷ <https://breast.predict.nhs.uk/>

¹⁰⁸ <http://www.lifemath.net/cancer/>

¹⁰⁹ <https://doi.org/10.1007/s10549-015-3490-4>

¹¹⁰ <https://doi.org/10.1007/s10549-019-05399-2>

¹¹¹ <https://iknl.nl/nieuws/2019/cancermath-geeft-nauwkeurige-voorspelling-overlevi>

acceptable prediction capabilities for the countries in which it was validated.^{112 113} No validation study for PREDICT in the Netherlands was found although a great many models were externally validated using IKNL data.¹¹⁴ However, most of these models are not directly publicly available. Of course, it goes without saying that the use of prediction tools needs to be done with extreme care, especially in subgroups.^{115 116}

The studies, of which some were reviews, show how the techniques used to predict breast cancer survival have changed across the years. From more ‘traditional’ forms of modelling (such as Logistic regression) to full-fledge machine learning models (e.g., gradient boosting, support vector machines), each model has its advantages and disadvantages. Some authors openly propose ensemble models which combine different models into one, but at a loss of being able to explain the underlying mode of action. Machine learning models in general are more difficult to explain as the coefficients that determine the direction and strength of a variable in connection with the outcome are replaced by a normalized value showing its importance on a 0% - 100% scale.

Results of our previous work

The result of previous work is described in [this Medium post](#) and discussed with IKNL in a separate meeting, long before we conducted this study. Hence, we will not recreate the visualizations of the findings we presented back then. However, we will highlight some of the findings we wrote, since they will affect our final steps in this endeavor. This is what we wrote, back then, after conducting several models via several methods:

“It’s decision time. Let’s make the model we want to have and focus a bit more. Based on the above, I will make the following choices:

1. *Females only → although breast cancer is a possibility in men, we have more data on females.*
2. *Tumor kind 501300 only (invasive mamma: MJ) → the majority of the data is situated here.*
3. *Will not include the inclusion year — this is a proxy for survival.*
4. *From the penalized (cox regression: MJ) models, only the following variables were deemed important: age, differential grade, tumor position, hormonal indicators, Mari-procedure, and tumor stage.*
5. *From the Random Forest Models, only the following variables were deemed important: age, Mari-procedure, tumor size, tumor stage, sentinel node procedure*

¹¹² <https://doi.org/10.1016/j.critrevonc.2020.102908>

¹¹³ <https://doi.org/10.1186/s13058-017-0852-3>

¹¹⁴ <https://doi.org/10.1016/j.breast.2023.04.003>

¹¹⁵ <https://doi.org/10.1016/j.jclinepi.2022.10.016>

¹¹⁶ <https://doi.org/10.1016/j.breast.2023.01.006>

performed, surgery performed, number of positive lymph nodes found, differential grade, and hormonal receptors.”

Results of our current work

Building a prediction model on the synthetic breast cancer data made available via the IKNL was of course not a novel exercise and builds strongly on the methods of our previous work. In short, we built a clinical prediction model, including patient, clinical, and treatment characteristics identified as important by the literature and our previous work. Included were *age, tumor stage, tumor grade, inclusion year, estrogen receptor, progesterone receptor, HER2 status, number of positive lymph nodes, tumor size, MARI procedure, sentinel node procedure, hormonal treatment, target treatment, chemotherapy, radiotherapy, and type of surgery.*

Below you can see the results from the Elastic Net Cox Regression which, based on cross-validated optimal L1 and L2 lambda values, returned 19 /44 coefficients as non-zero (Figure 109). A separate endeavor, using a different package, led to a more drastic reduction of variables (Figure 110 and text with a black background after that). Here, only *age, estrogen receptor, MARI procedure and chemotherapy* remained. However, this did not strike us as reliable, since almost all models found in the literature indicate some meaning towards tumor stage, grade and positive lymph nodes found.

```
Penalized cox regression object
44 regression coefficients of which 19 are non-zero
```

```
Loglikelihood = -17927.43
L1 penalty = 52.20724      at lambda1 = 44.12653
L2 penalty = 57.62374      at lambda2 = 206.4989
```

	variables	coefficient	include
1	Stage1	-0.013237228	yes
2	Stage2	0.000000000	no
3	Stage3	0.000000000	no
4	Stage4	0.000000000	no
5	StageM	0.030637796	yes
6	chemoNo	0.019310487	yes
7	chemoPost-surgery	0.000000000	no
8	chemoPre-and-post-surgery	-0.096299035	yes
9	chemoPre-surgery	-0.010879290	yes
10	chemoYes, no surgery	0.684802265	yes
11	diffgradHigh grade	0.000000000	no
12	diffgradIntermediate grade	0.024092056	yes
13	diffgradLow grade	0.000000000	no
14	er_stat_fact1	0.042385212	yes
15	her2_stat_fact1	0.000000000	no
16	hormNo	0.000000000	no
17	hormPost-surgery	0.000000000	no
18	hormPre-and-post-surgery	0.000000000	no
19	hormPre-surgery	0.000000000	no
20	hormYes, no surgery	0.000000000	no
21	incjr	0.000000000	no
22	mariYes	0.662064804	yes
23	ns(leeft, 3)1	1.245694884	yes

```

24 ns(leeft, 3)2 -2.234997078 yes
25 ns(leeft, 3)3 2.000653809 yes
26 ns(pos_lymf, 3)1 0.000000000 no
27 ns(pos_lymf, 3)2 0.000000000 no
28 ns(pos_lymf, 3)3 0.000000000 no
29 ns(tum_afm, 3)1 0.000000000 no
30 ns(tum_afm, 3)2 0.009238974 yes
31 ns(tum_afm, 3)3 -0.046820908 yes
32 pr_stat_fact1 0.000000000 no
33 rtNo 0.000000000 no
34 rtPost-surgery 0.000000000 no
35 rtPre-surgery 0.000000000 no
36 swkNo 0.000000000 no
37 swkNot registered -0.096391501 yes
38 swkYes 0.042664454 yes
39 uitgebr_chir_codeAblatio (no OKD) 0.000000000 no
40 uitgebr_chir_codeAmputation (with OKD) 0.000000000 no
41 uitgebr_chir_codeLumpectomie (no OKD) 0.031359364 yes
42 uitgebr_chir_codeLumpectomie (with OKD) 0.000000000 no
43 uitgebr_chir_codeLymph node dissection 0.345372519 yes
44 uitgebr_chir_codeResection other (coincidence) -0.043812598 yes

```

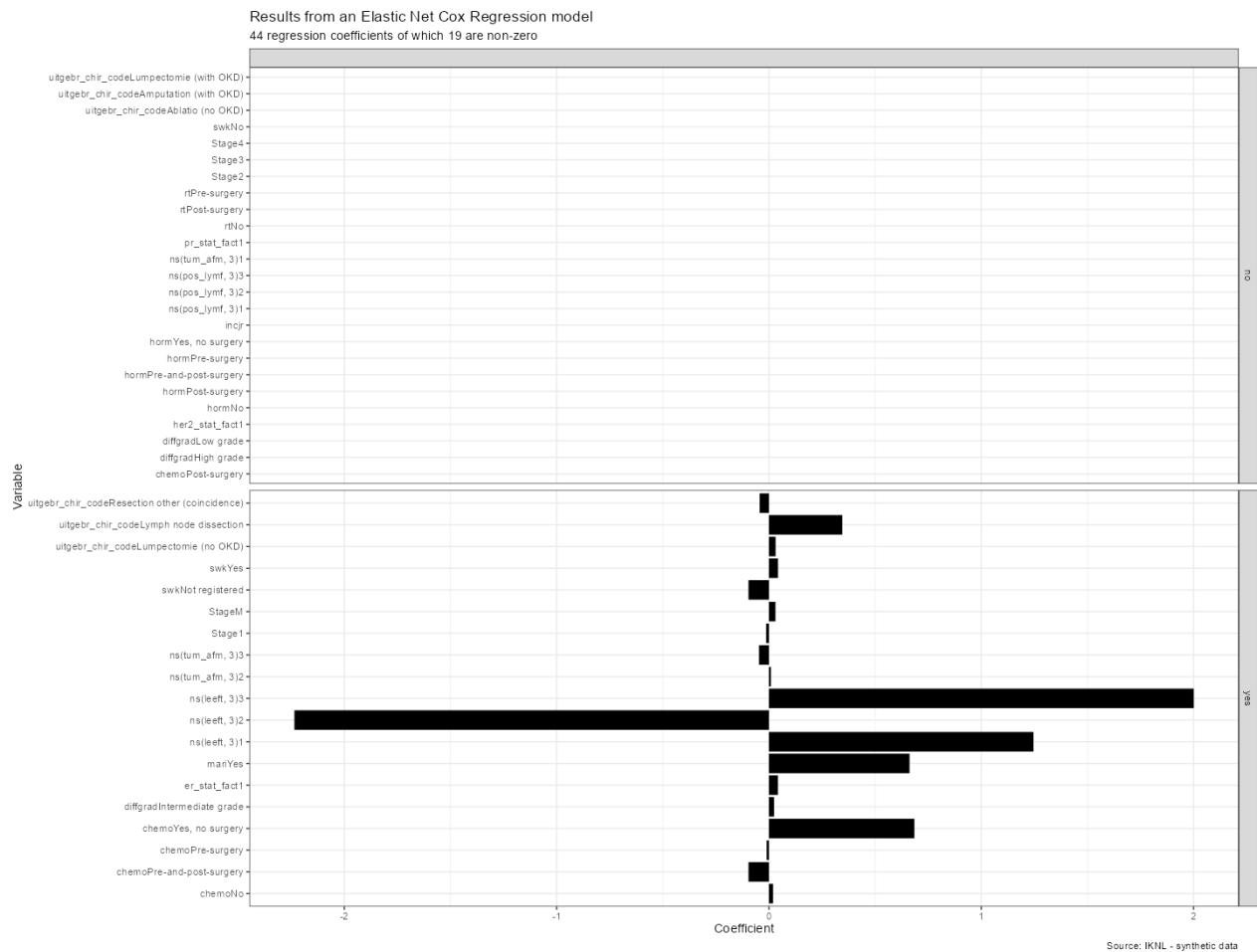


Figure 109. Results from an Elastic Net Cox Regression model showing which predictors should be included or not (penalized package)

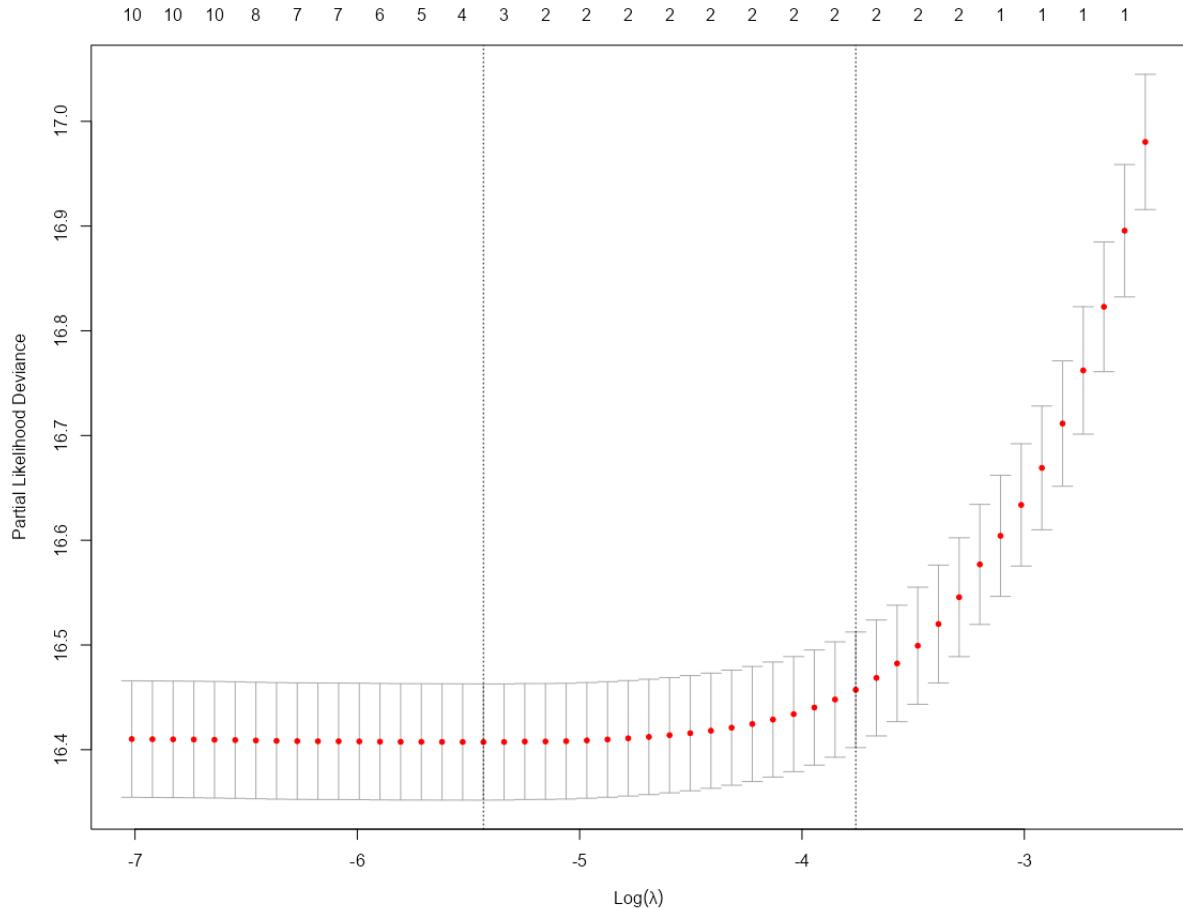


Figure 110. Cross-validation results from the GLMNET Cox model.

```
15 x 1 sparse Matrix of class "dgCMatrix"
 1
leeft           4.948130e-02
incjr           .
diffgrad        .
Stage            .
er_stat_fact   3.044389e-02
pr_stat_fact   .
her2_stat_fact .
tum_afm         .
pos_lymf        .
swk              .
mari             6.763193e-01
horm              .
chemo            -1.515464e-05
rt               .
uitgebr_chir_code .
```

In the end, we decided to stick to a full clinical model keeping all 44 variables. This of course goes against the basic sense of integrating L1 and L2 regularization, but we at least want to

see how such a full clinical model would perform. In the next four figures (Figure 111 to Figure 114), you can see the bootstrapped calibration plots for predicting the probability of surviving 500, 1000, 1500, and 2000 days following diagnosis.

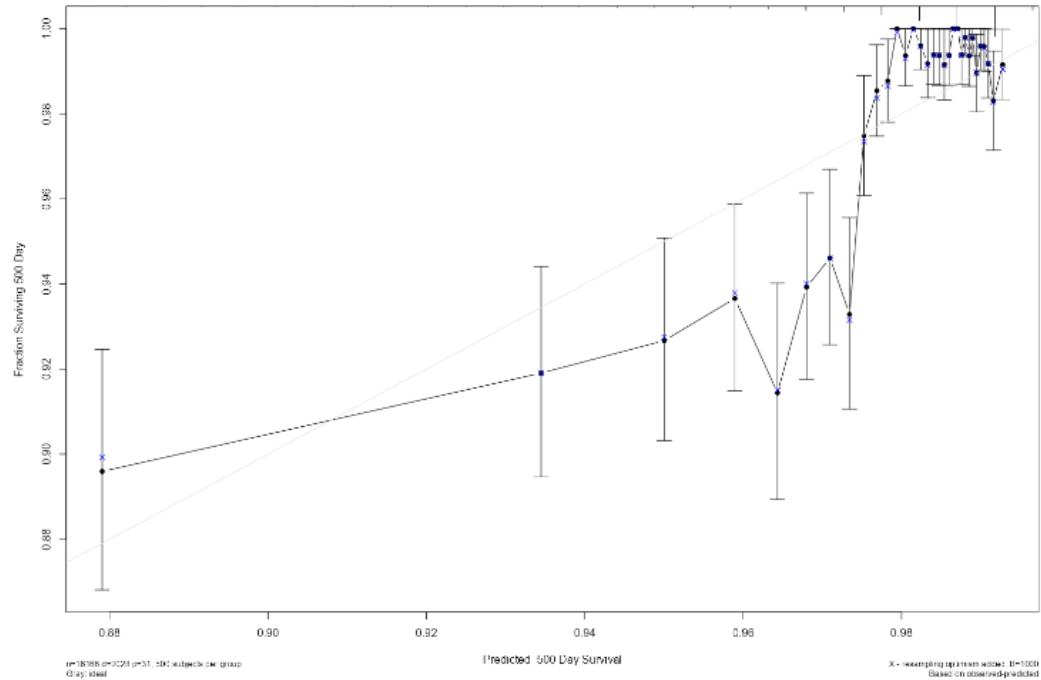


Figure 111. Calibration plot for 500 day survival of a Cox Regression model based on 1000 bootstrap samples.

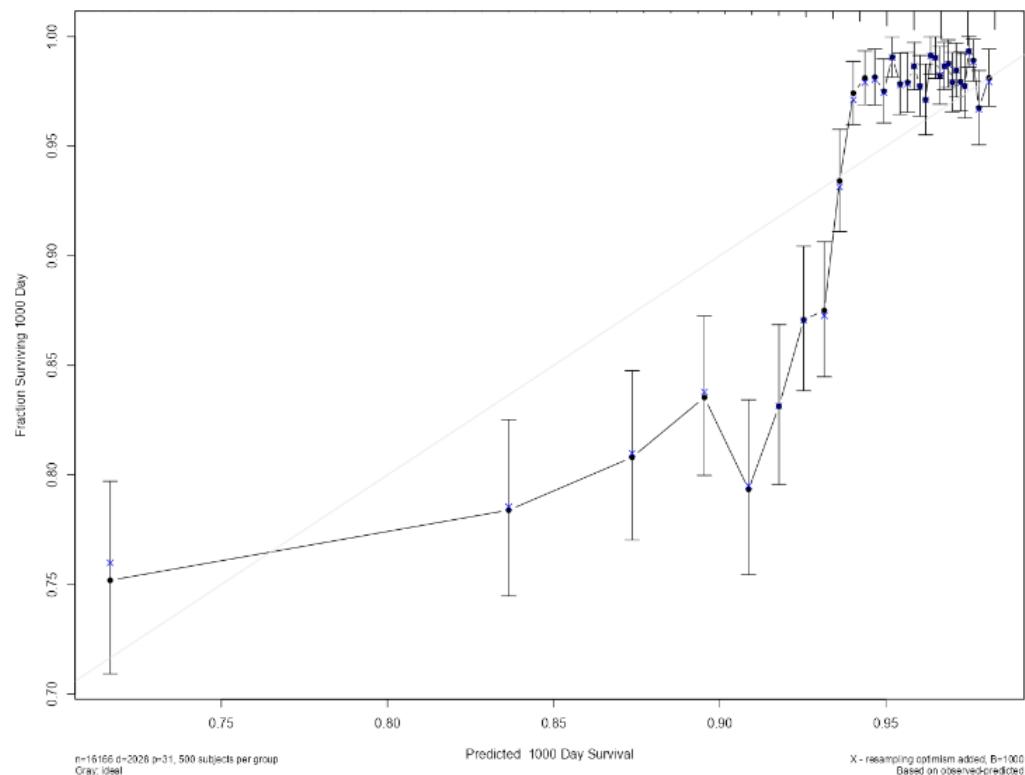


Figure 112. Calibration plot for 1000 day survival of a Cox Regression model based on 1000 bootstrap samples.

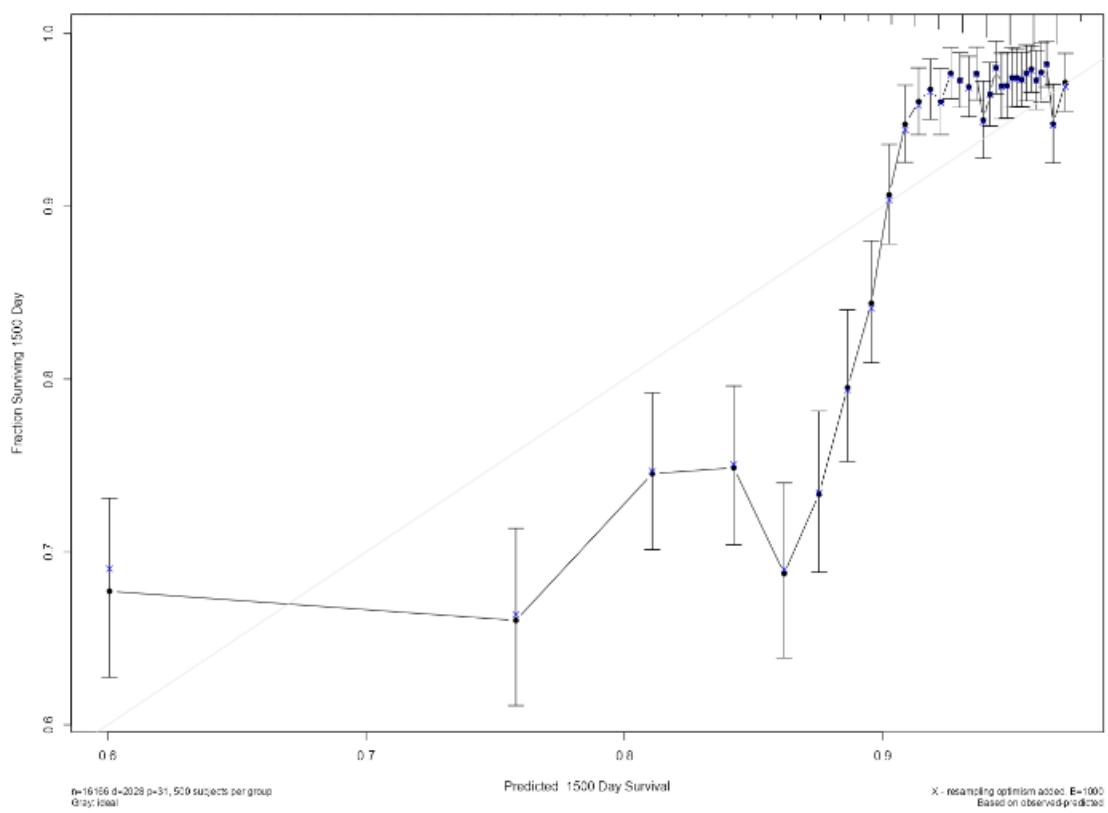


Figure 113. Calibration plot for 1500 day survival of a Cox Regression model based on 1000 bootstrap samples.

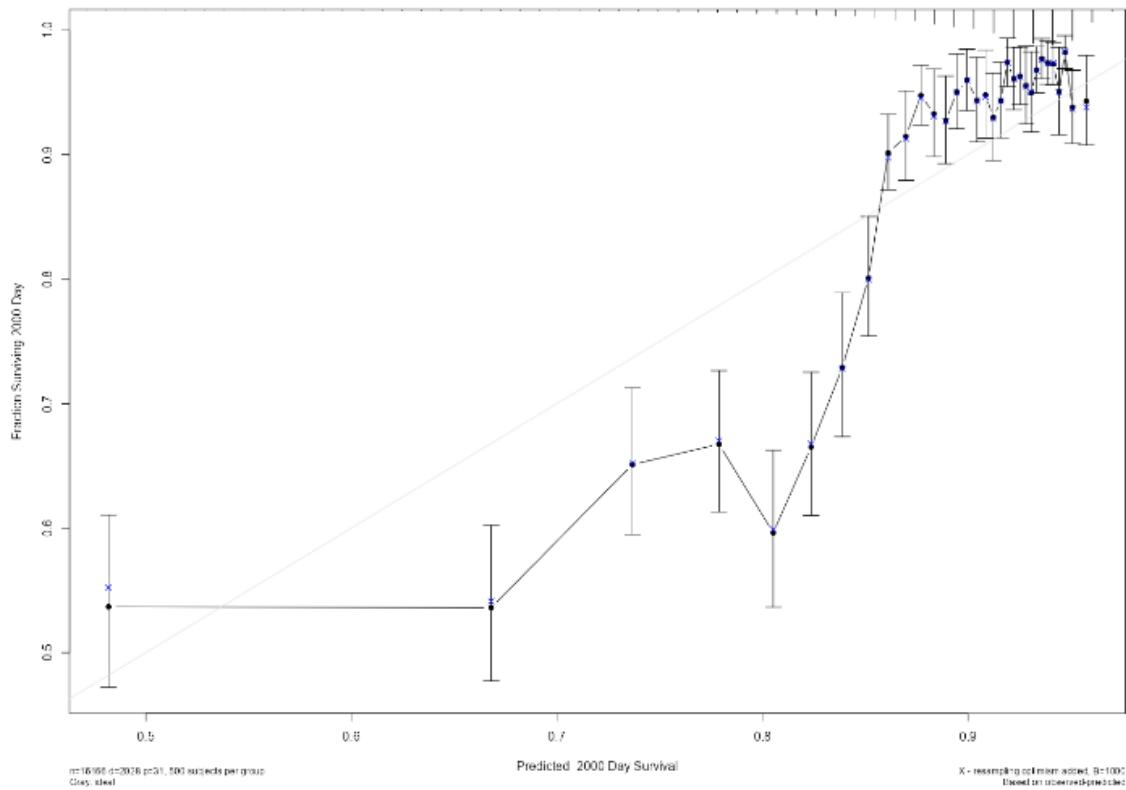


Figure 114. Calibration plot for 2000 day survival of a Cox Regression model based on 1000 bootstrap samples.

The plots above are not that bad, but they are not good either. The further in time we want to predict the worse the model gets to the point where it is seriously underestimating survival probability. However, for our purpose, only the 500- and 1000-day plots are useful, and one could say that even the 1000 day plot exceeds our purpose of looking at the first two years following diagnosis. Nevertheless, these plots as well show some underestimation of the survival probability.

Next, we look at various residual plots for misbehaving residuals that could indicate bad model fit and/or bad model selection. At first sight, the model seems to be doing fine (Figure 115), although there seems to be some sort of two-class structure underneath the data. By this we mean that the Deviance (Figure 116) and Martingale residuals (Figure 117) split into two clouds showing no even spreads for certain values of the linear predictor.

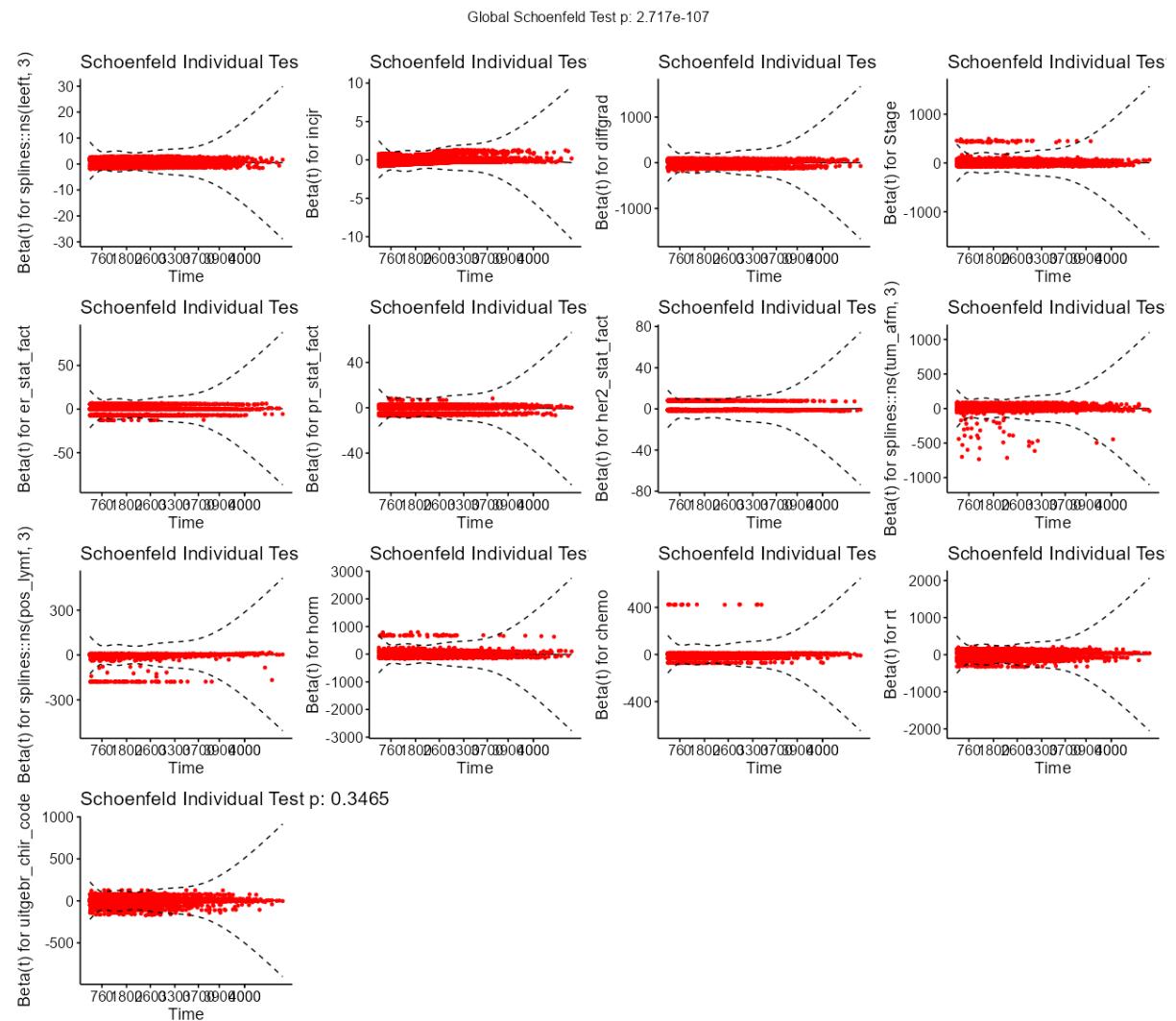


Figure 115. Schoenfeld residuals for Cox Regression model.

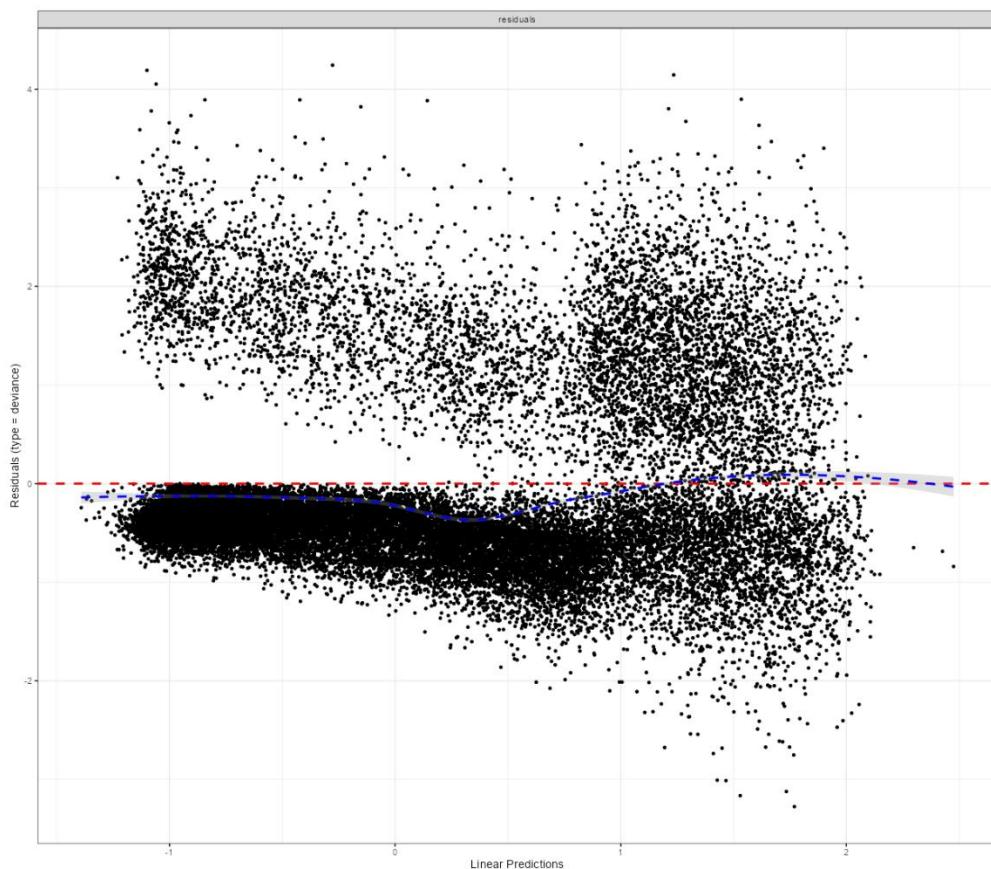


Figure 116. Deviance residuals coming from Cox Regression model.

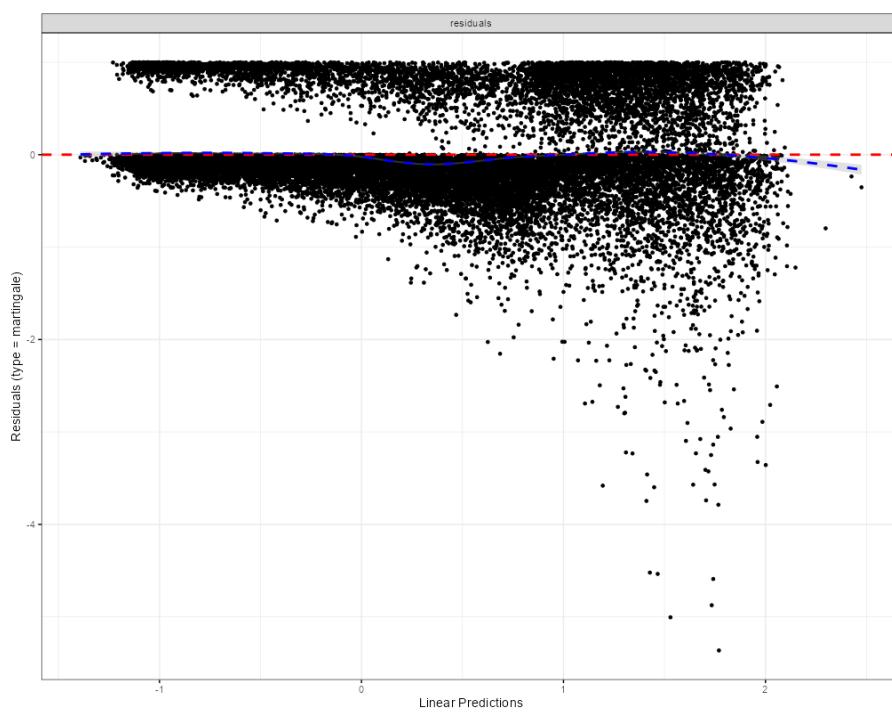


Figure 117. Martingale residuals coming from Cox Regression model.

Below you can see the output of the model in a table form (Figure 118) and a forest plot (Figure 119). The forest plot is always quite informative to use when trying to investigate the strength of a coefficient to its variation (i.e., the standard error). This actually seems to be quite reasonable which should not surprise that much considering the size of the dataset (between 30k and 45k of data were used for the models).

Characteristic	HR	95% CI	p-value
<code>splines:ns(left, 3)</code>			
<code>splines:ns(left, 3)1</code>	6.23	5.43, 7.20	<0.001
<code>splines:ns(left, 3)2</code>	0.16	0.10, 0.26	<0.001
<code>splines:ns(left, 3)3</code>	13.4	11.5, 15.7	<0.001
<code>incr</code>	1.02	1.01, 1.03	<0.001
<code>difgrad</code>			
Low grade	—	—	
Intermediate grade	1.01	0.95, 1.07	0.9
High grade	0.97	0.90, 1.05	0.4
Stage			
1	—	—	
2	1.03	0.97, 1.09	0.3
3	1.03	0.94, 1.12	0.6
4	1.15	0.88, 1.50	0.3
M	1.06	0.82, 1.38	0.6
<code>er_stat_fact</code>			
0	—	—	
1	1.01	0.92, 1.10	0.9
<code>pr_stat_fact</code>			
0	—	—	
1	1.02	0.98, 1.12	0.2
<code>her2_stat_fact</code>			
0	—	—	
1	0.98	0.91, 1.05	0.5
<code>splines:ns(tum, 3)</code>			
<code>splines:ns(tum, 3)1</code>	1.17	0.96, 1.42	0.13
<code>splines:ns(tum, 3)2</code>	0.84	0.65, 1.09	0.2
<code>splines:ns(tum, 3)3</code>	0.69	0.43, 1.11	0.12

Figure 118. Summary table showing hazard ratios for each variable included in the Cox Regression model (click on link for html page).

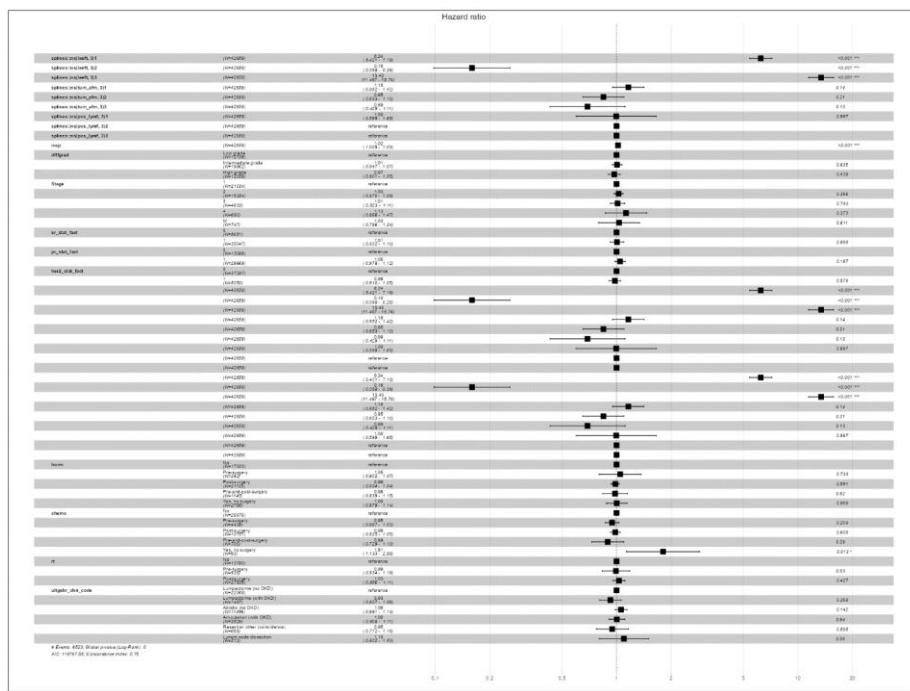


Figure 119. Forest plot of coefficients included in Cox Regression model.

Nevertheless, we decide to move on and look more closely at the model and its ability to differentiate if we would offer it multiple different “patients” (Figure 120). Figure 121 shows what happens to the survival curve if we enter these three patients to the Cox Regression model we made. It seems that the model can differentiate, especially as time moves on.

However, for the first two years it is basically the same according to this model. The rest of the graphs (Figure 122 and Figure 123) show the same. Especially Figure 123 shows that the survival probabilities for the first years are different per curve.

	leeft	incjr	diffgrad	Stage	er_stat_fact	pr_stat_fact	her2_stat_fact	tum_afm	pos_lymf	norm	chemo	rt	target	uitgebr_chir_code	ID
1	30	2019	Intermediate grade	1	1	1	0	15	0	Post-surgery	No	No	No	Lumpectomie (with OKD)	1
2	40	2019	Intermediate grade	2	1	1	0	20	3	Post-surgery	No	Post-surgery	No	Lumpectomie (with OKD)	2
3	59	2019	Intermediate grade	2	1	1	1	25	6	Post-surgery	No	Post-surgery	No	Lumpectomie (with OKD)	3

Figure 120. New data frame for making predictions based on fitted Cox Regression model – included are three fictitious patients.

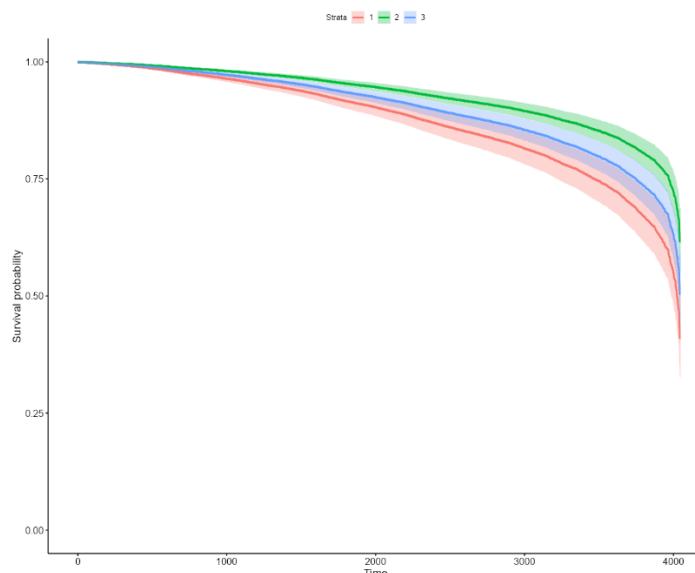


Figure 121. Visualization of survival probabilities for each the three patients coming from the new data frame.

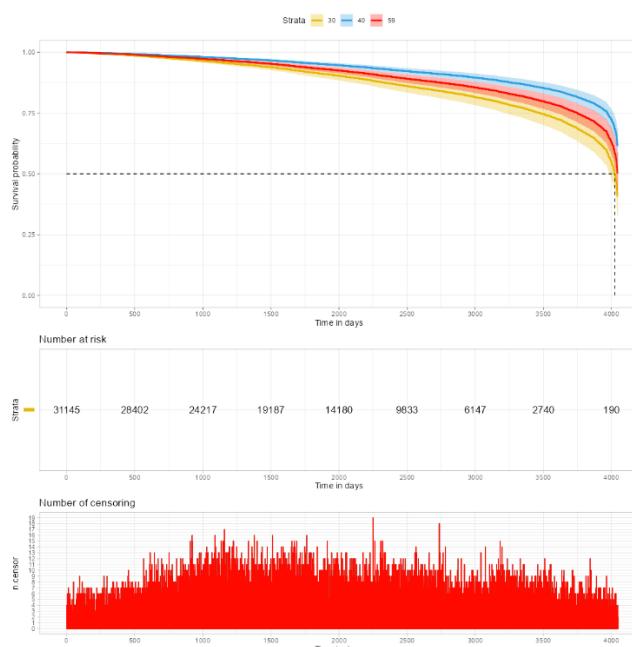


Figure 122. More complicated picture on survival probabilities for the three patients.

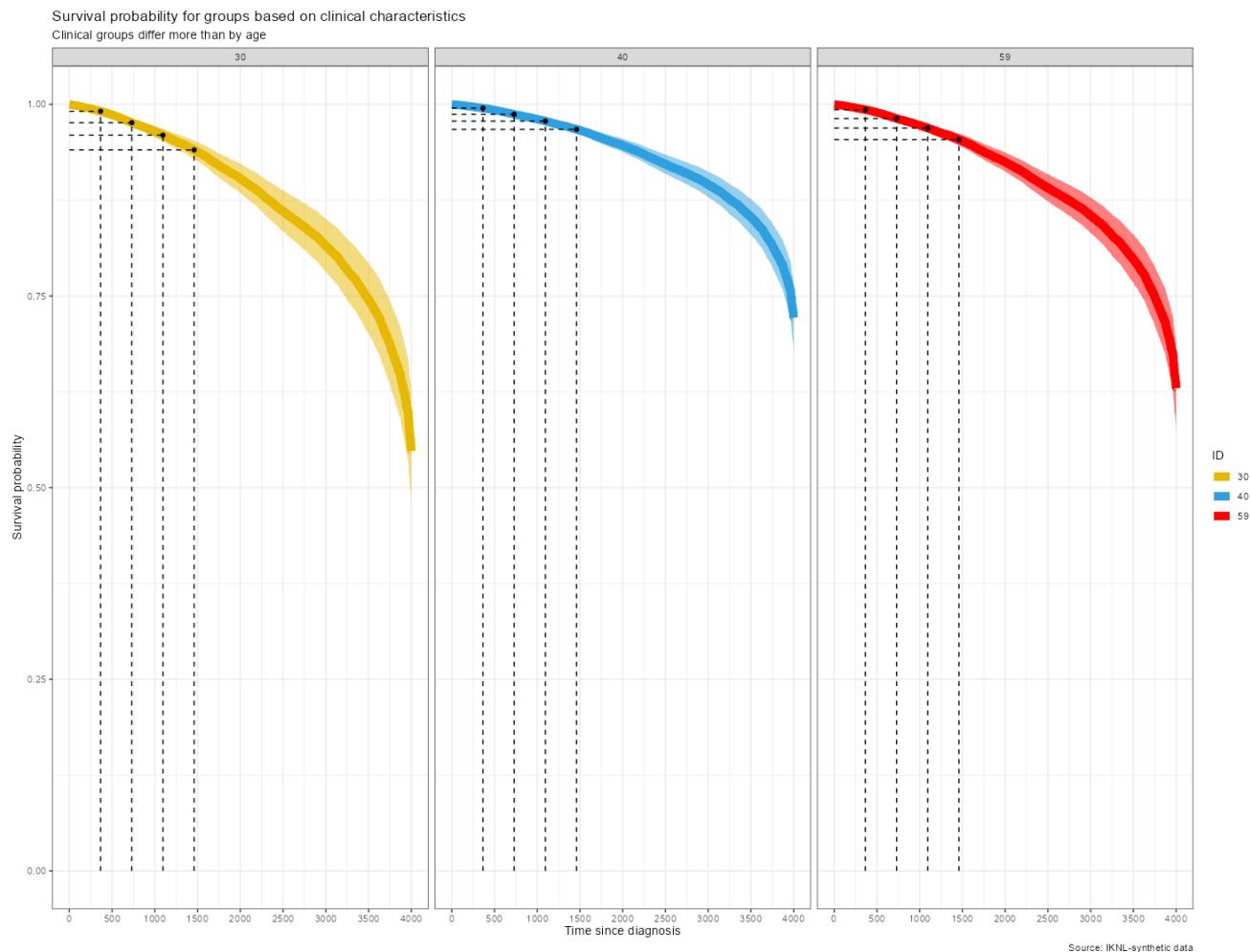


Figure 123. Survival probabilities for three patients showing the survival probability for 1,2,3 and 4 years.

Last but not least, we want to create a nomogram, which is basically a nice interactive way of finding survival probabilities for each possible combination of variables included (also ones that do not make sense clinically but in the model world anything goes). However, to do so, we needed to remove all missing data, and we decided to delete the MARI procedure and the sentinel procedure, since neither PREDICT nor CancerMath has anything included about them. Before creating the model on a subset of the data, we checked the concordance statistics which were quite similar. We will later show why we do not believe that these will easily change anyhow, for better or worse.

```
> concordance(fit_sub)
Call:
concordance.coxph(object = fit_sub)

n= 31145
Concordance= 0.7487 se= 0.003554
concordant discordant tied.x      tied.y      tied.xy
   83526242    28042261       27       5646        0
> concordance(fit)
Call:
concordance.coxph(object = fit)

n= 31145
```

```

Concordance= 0.7488 se= 0.003555
concordant discordant tied.x tied.y tied.xy
 83544161    28024346      23      5646      0

```

Figures 124, 125, and 126 show some of the parts of the dynamic nomogram which can produce survival plots and predicted survival points based on the variable included. Below Figure 126 you can see the predicted survival probability at the median age (1848 days) for the three patients shown in Figure 120.

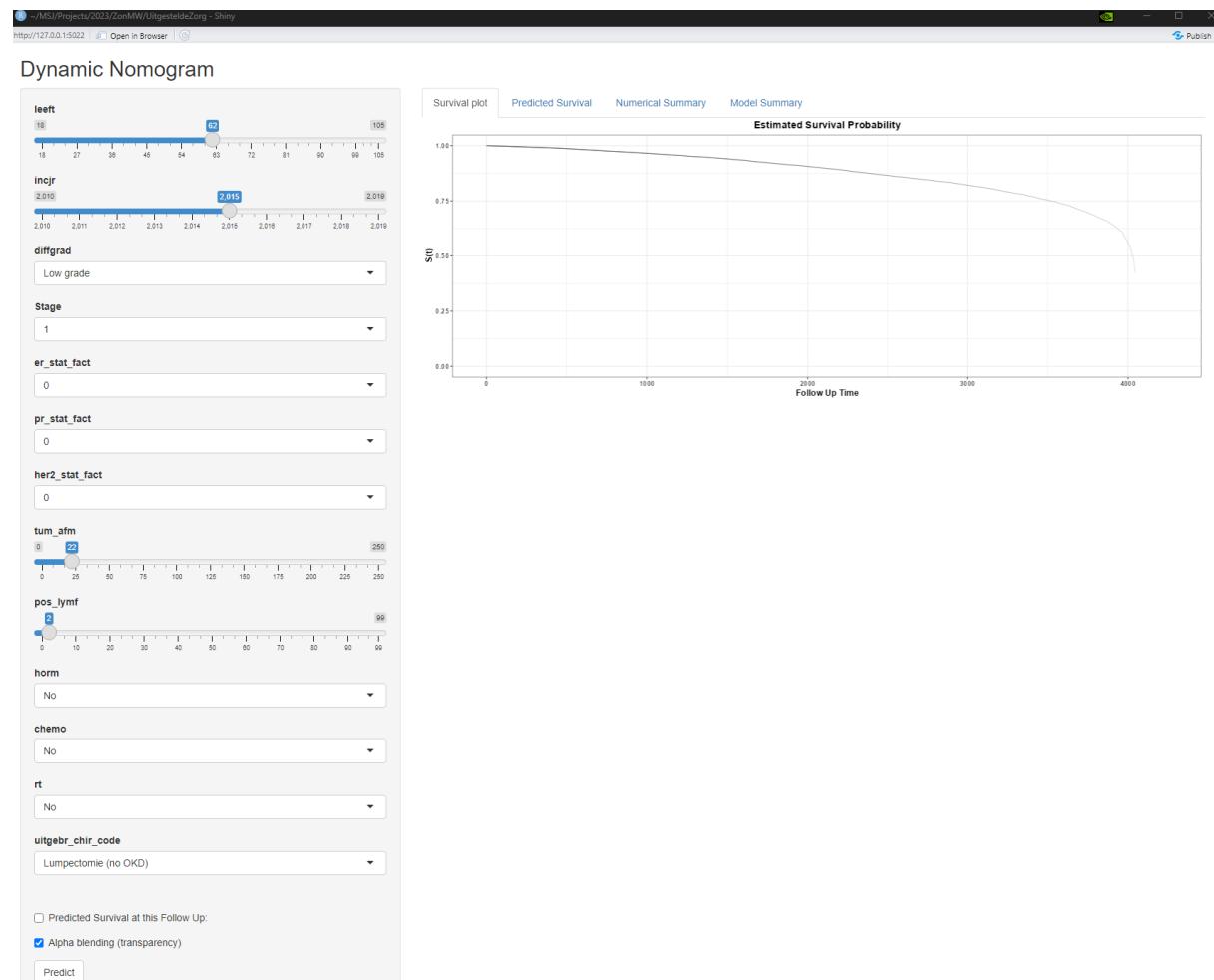


Figure 124. Picture of a dynamic nomogram based on the Cox Regression model.

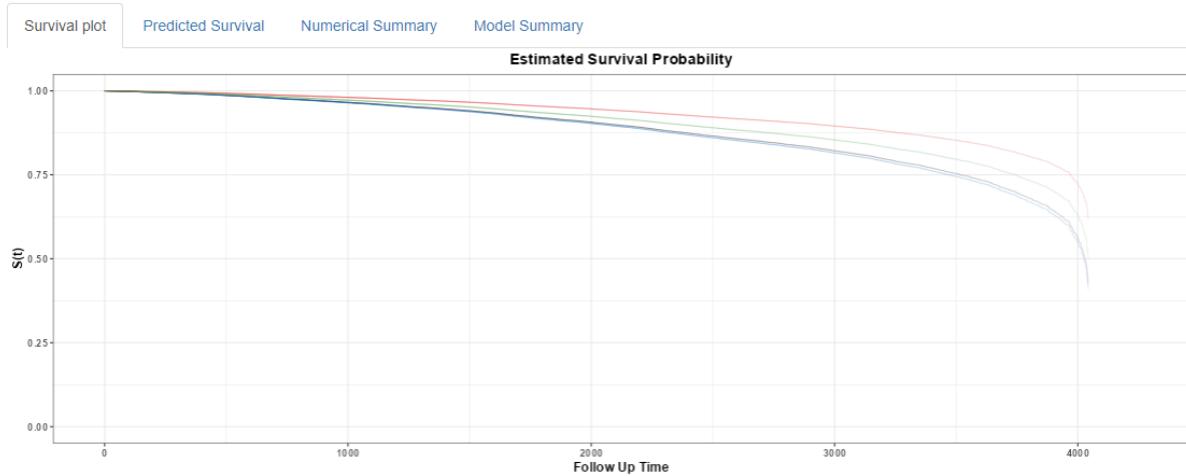


Figure 125. Survival plot as shown in a dynamic nomogram.

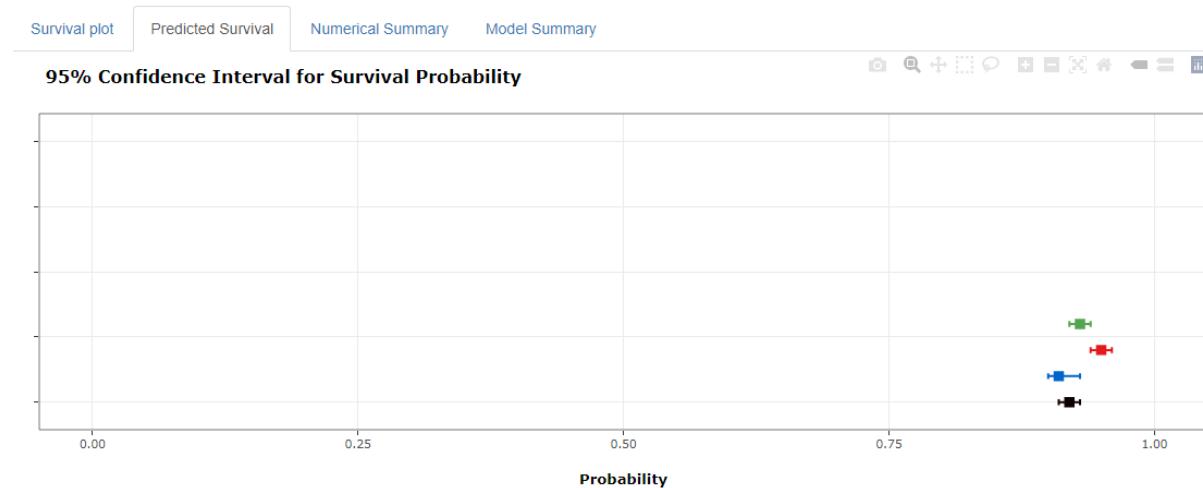


Figure 126. Predicted Survival as shown in a dynamic nomogram.

	vit_stat_int	leeft_injr	diffgrad	Stage	er_stat_fact	pr_stat_fact	her2_stat_fact	tum_afm	pos_lymf	horm	chemo	rt	uitgebr_chir_code	Prediction	Lower.bound	Upper.bound	
1	1,840	62	2,015	Low grade	1	0	0	0	22	2	No	No	No	Lumpectomie (no OKD)	0.920	0.910	0.930
2	1,840	30	2,019	Intermediate grade	1	1	1	0	15	0	Post-surgery	No	No	Lumpectomie (with OKD)	0.910	0.900	0.930
3	1,840	40	2,019	Intermediate grade	2	1	1	0	20	3	Post-surgery	No	Post-surgery	Lumpectomie (with OKD)	0.950	0.940	0.960
4	1,840	59	2,019	Intermediate grade	2	1	1	1	25	6	Post-surgery	No	Post-surgery	Lumpectomie (with OKD)	0.930	0.920	0.940

Validity of the IKNL Synthetic data

At the end of the exercise above, we once again started to doubt heavily about the applicability of the model and the data. The reason being that the Survival probabilities of the different patients (we included many more which we do not show here) all seem to have pretty good survival. Hence, we want to do a simple check in which we calculated the

median survival per inclusion year, tumor stage, and patient status (alive or dead). A graphical representation can be seen in Figure 127.

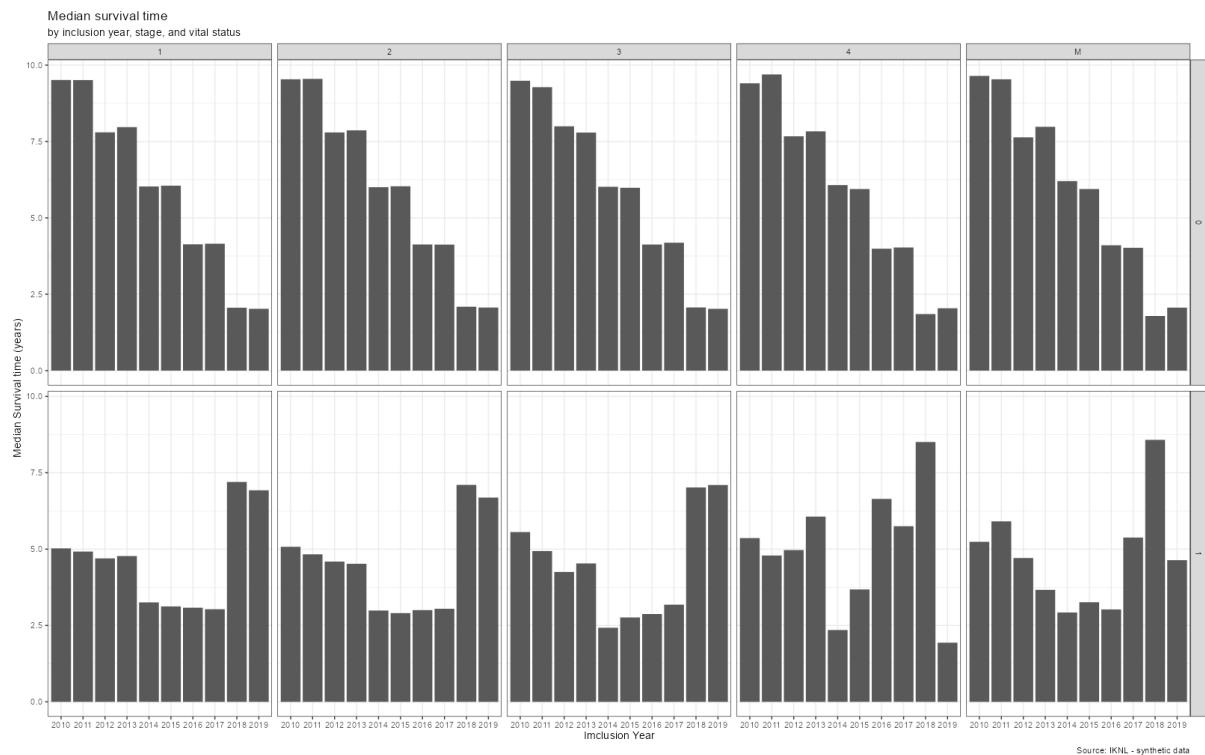


Figure 127. Bar chart of the median survival of the complete synthetic dataset split by inclusion year, tumor stage, and vital status.

However, this graph is not the best way of showing the expected changes in survival distributions, so we made Figure 128 to Figure 133 which shows the events across time. What is clear to see is that the curves mimic each other heavily although they become more raggedy as the tumor stage progresses. This is not in line with the survival probability per tumor stage as depicted on the website of IKNL in which a clear distinction can be made between the four stages¹¹⁷. All of this strengthens our belief that this dataset is not useful. Nevertheless, we did compare it with the other prediction models [here](#).

¹¹⁷ <https://iknl.nl/kankersoorten/borstkanker/registratie/overleving>

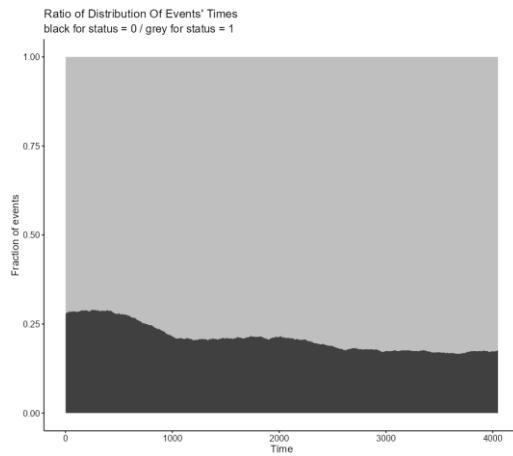


Figure 128. Ratio of distributions of Events / Time for all data.

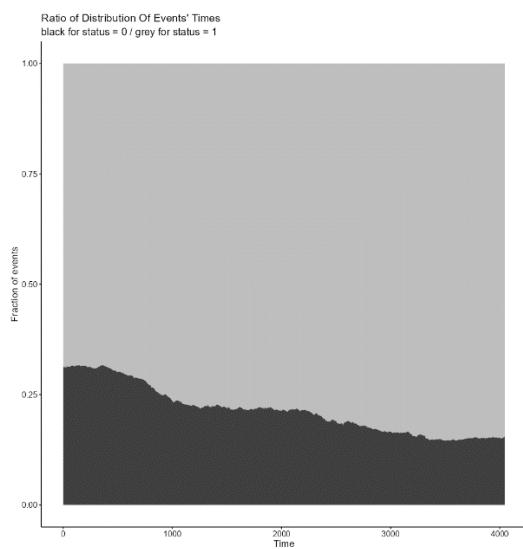


Figure 129. Ratio of distributions of Events / Time for Stage = 1.

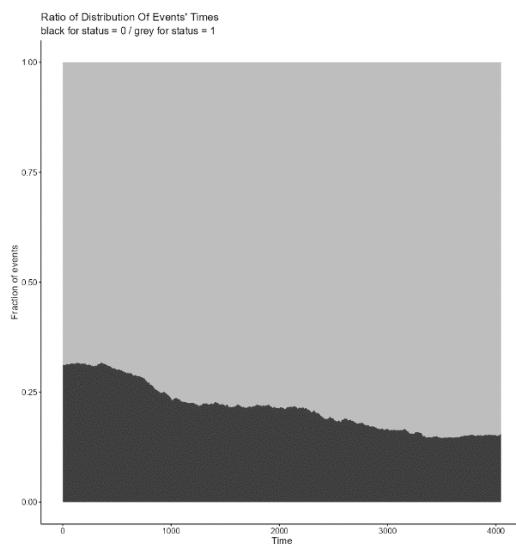


Figure 130. Ratio of distributions of Events / Time for Stage = 2.

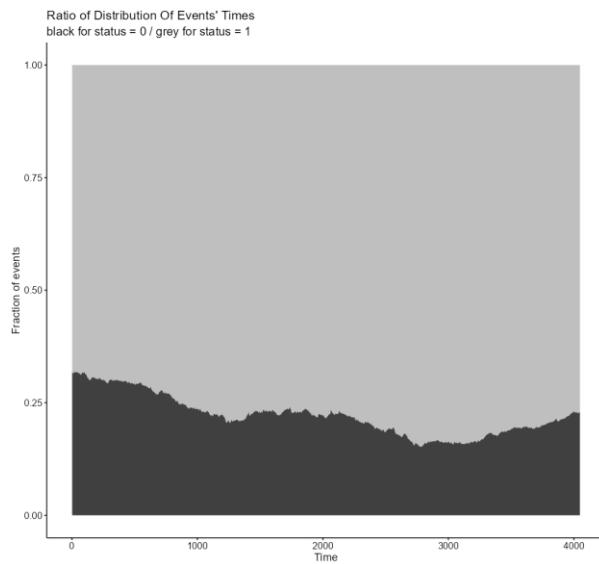


Figure 131. Ratio of distributions of Events / Time for Stage = 3.

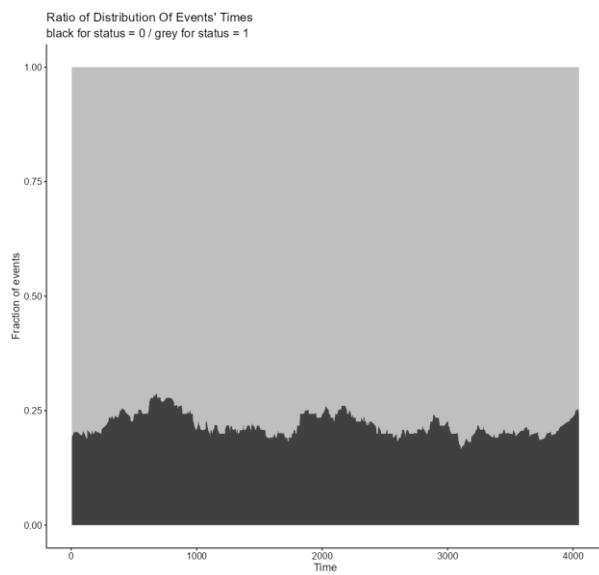


Figure 132. Ratio of distributions of Events / Time for Stage = 4.

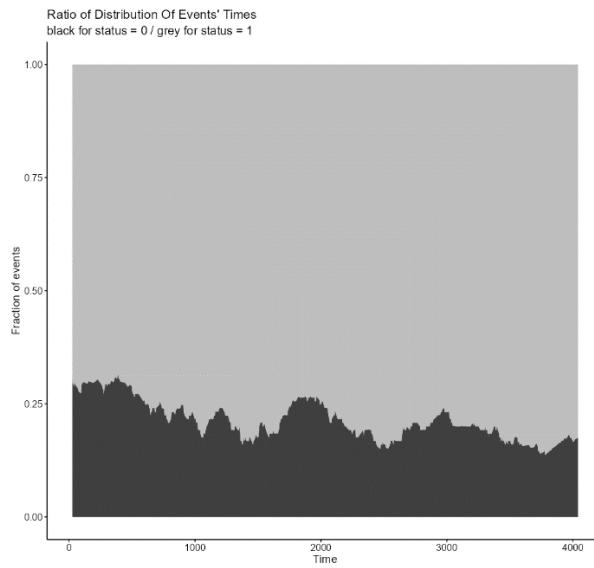


Figure 133. Ratio of distributions of Events / Time for Stage = M.

Most common groups in breast cancer IKNL synthetic data

Although we know that the descriptives coming from the synthetic dataset are limited, we do want to use it to assess if clinical groups shifted across time. By this we mean that we wanted to see if some years showed a trend reversal in the distribution of tumor stages, grades or treatment received. Despite synthetic data having somewhat limited descriptive potential it should still carry the underlying correlation / covariance matrix of the ‘real’ dataset.

Figures 134, 135 and 138, show tumor stages and tumor grades from 2010 to 2019, which are stable in their distribution according to the synthetic data of the IKNL.

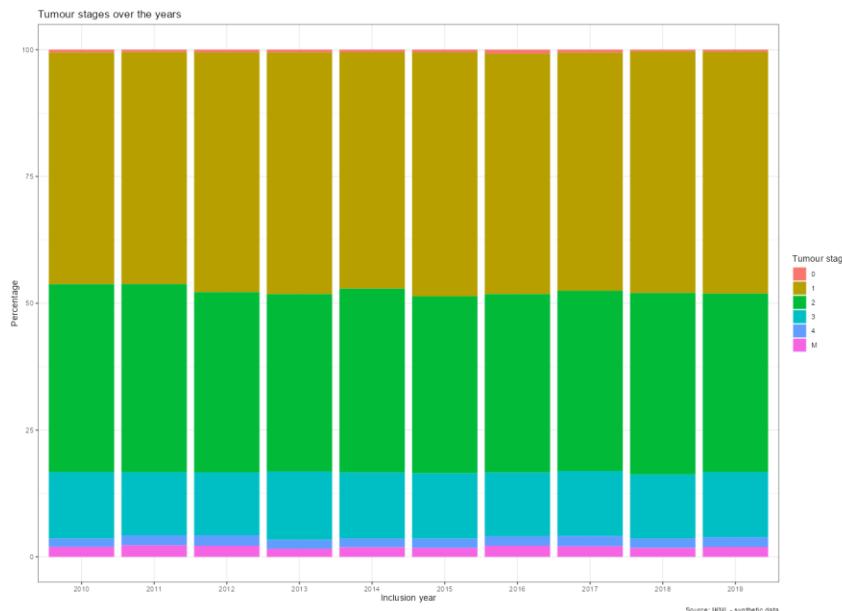


Figure 134. Distribution of tumor stages per inclusion year for breast cancer.

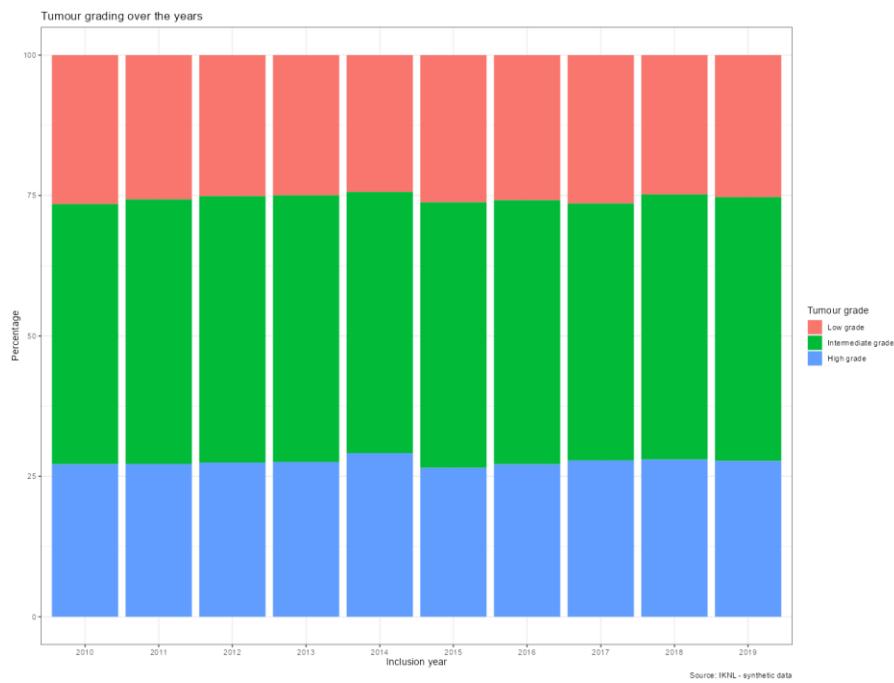


Figure 135. Distribution of tumor grading per inclusion year for breast cancer.

Plotting the distribution of genetically predisposed groups also does not show clear trend reversals across the years (Figures 136 and 137).

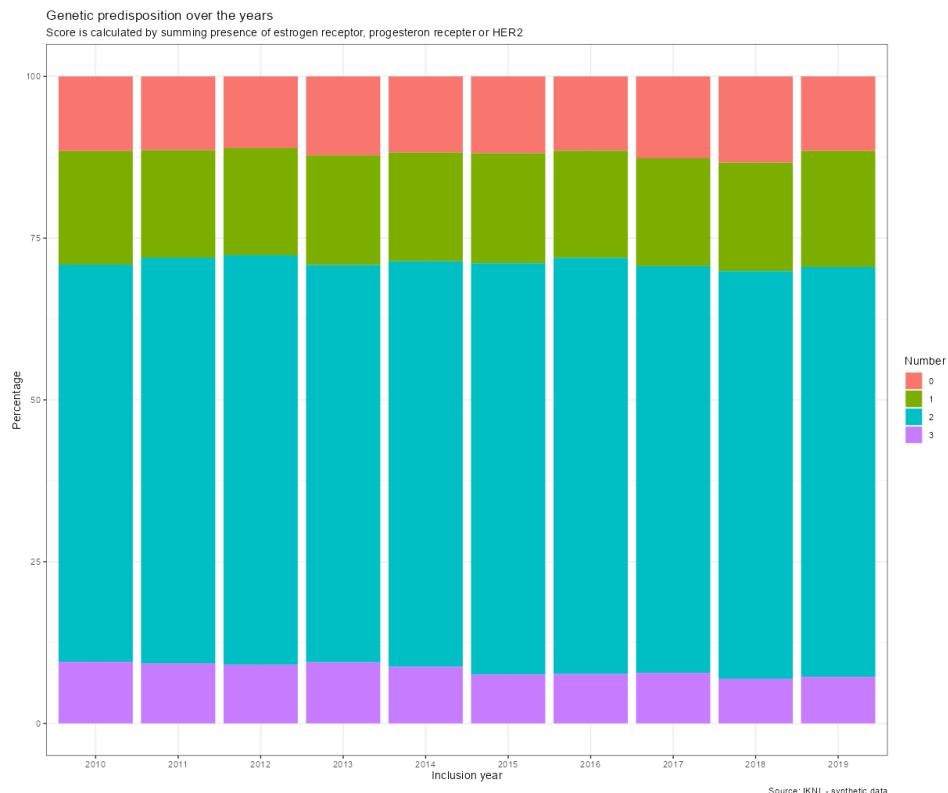


Figure 136. Distribution of genetic predisposition (HER2/Estrogen/Progesteron) per inclusion year for breast cancer.



Figure 137. Distribution of genetic predisposition (HER2; Estrogen; Progesteron) per inclusion year for breast cancer.

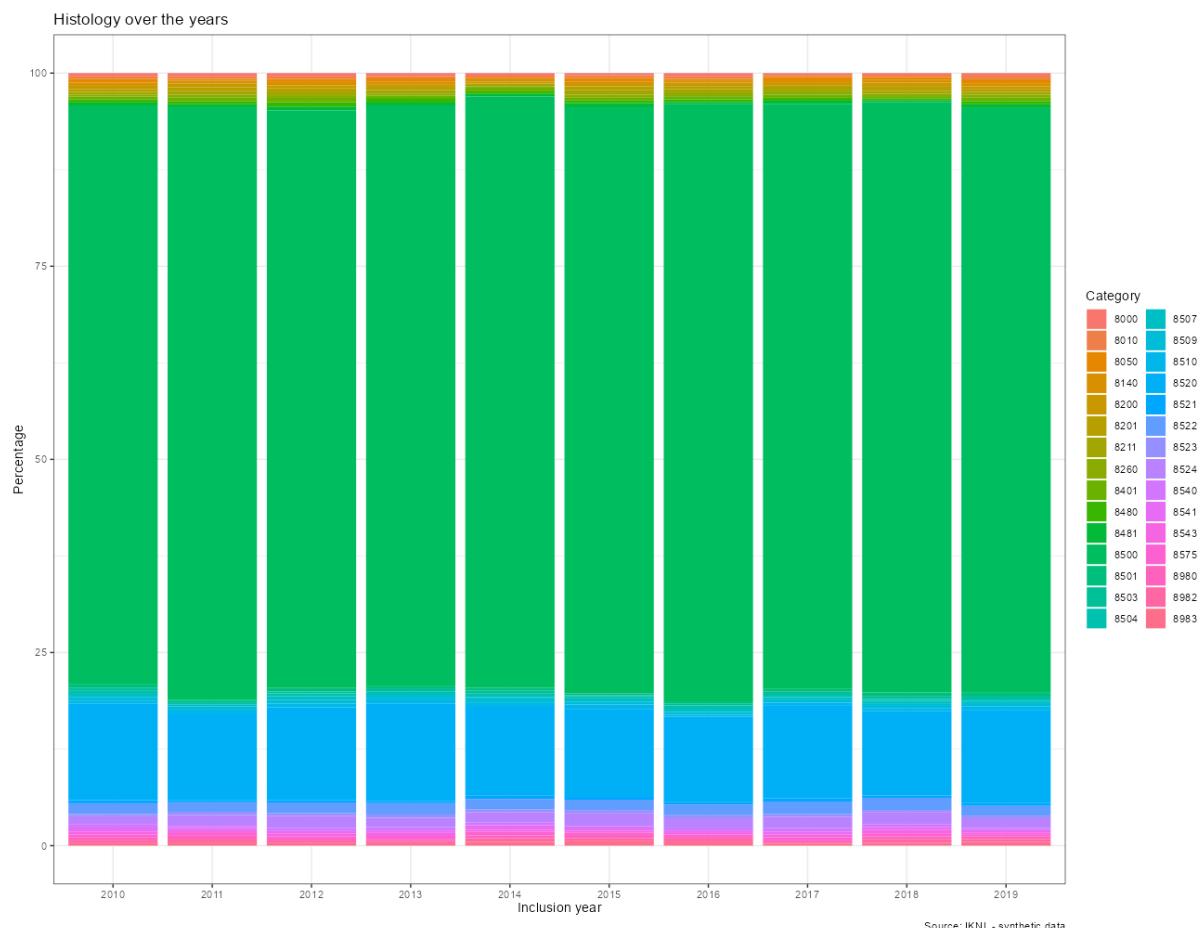


Figure 138. Distribution of histology per inclusion year for breast cancer.

Figures 139 - 141 do hint at a decreasing median age at time of inclusion, regardless of tumor stage or tumor grading. The rest of the figures, from 118 to 129, do not really show any trend reversals. It is not clear if this is a fluke of synthetic data, or reality.

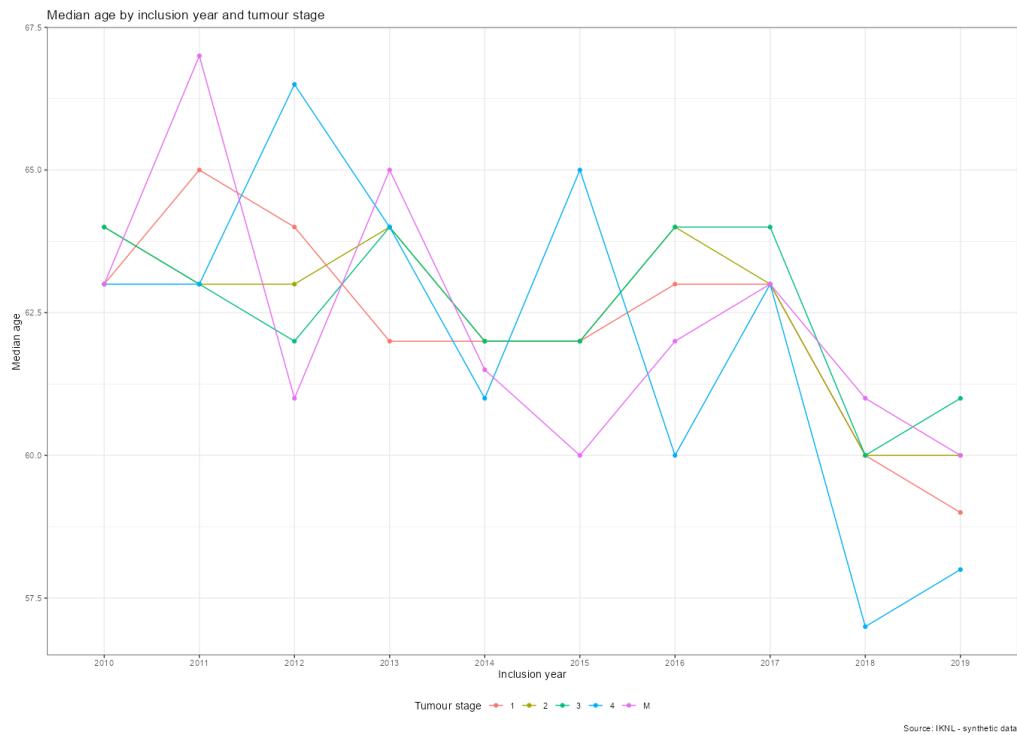


Figure 139. Median age per inclusion year and tumor stage for breast cancer.

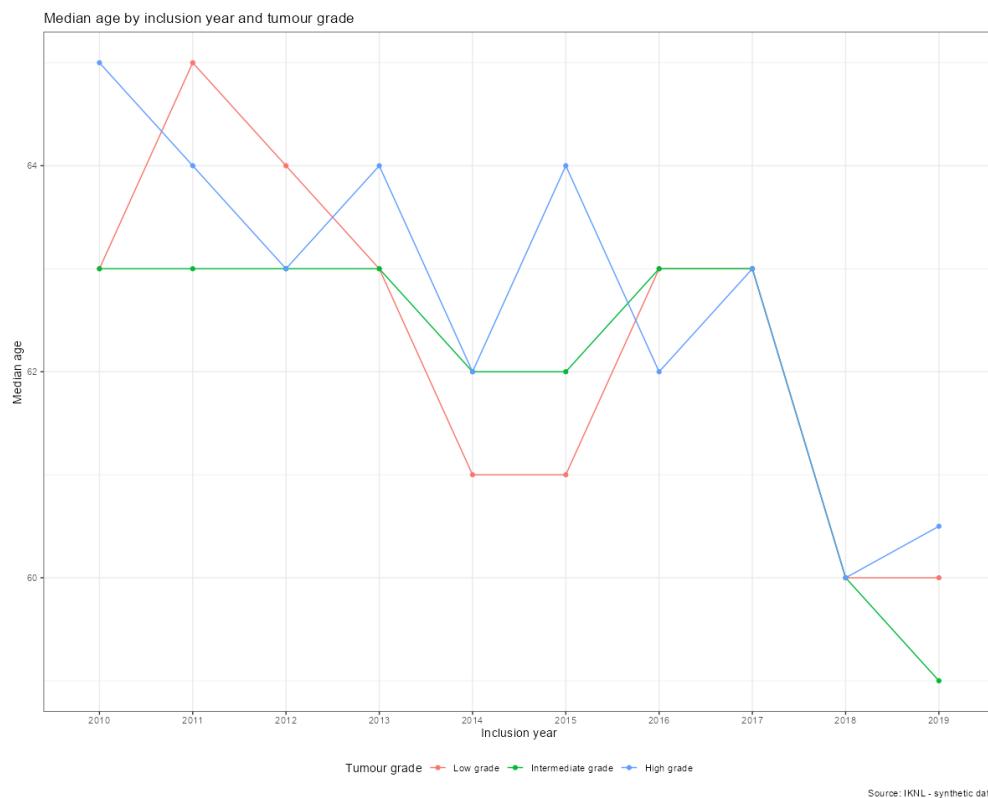


Figure 140. Median age per tumor grade per inclusion year.

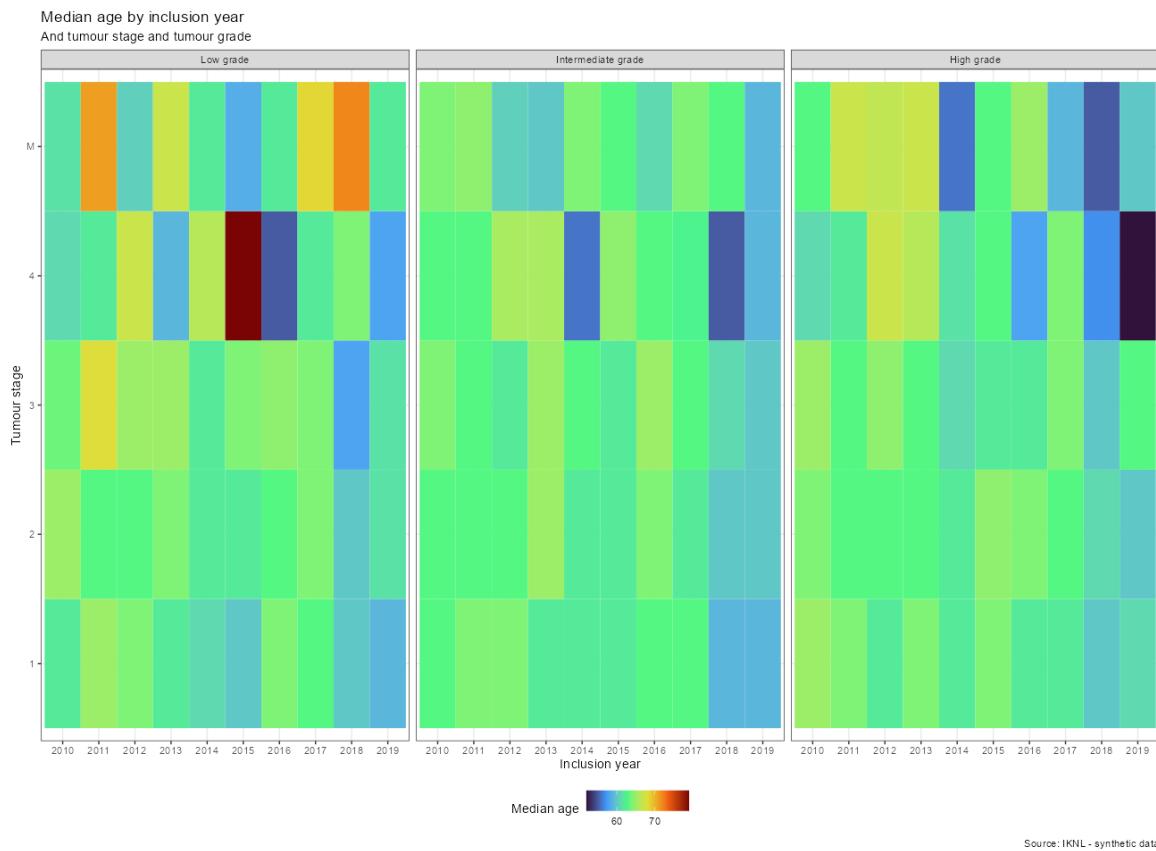


Figure 141. Median age per tumor stage, tumor grade and inclusion year.

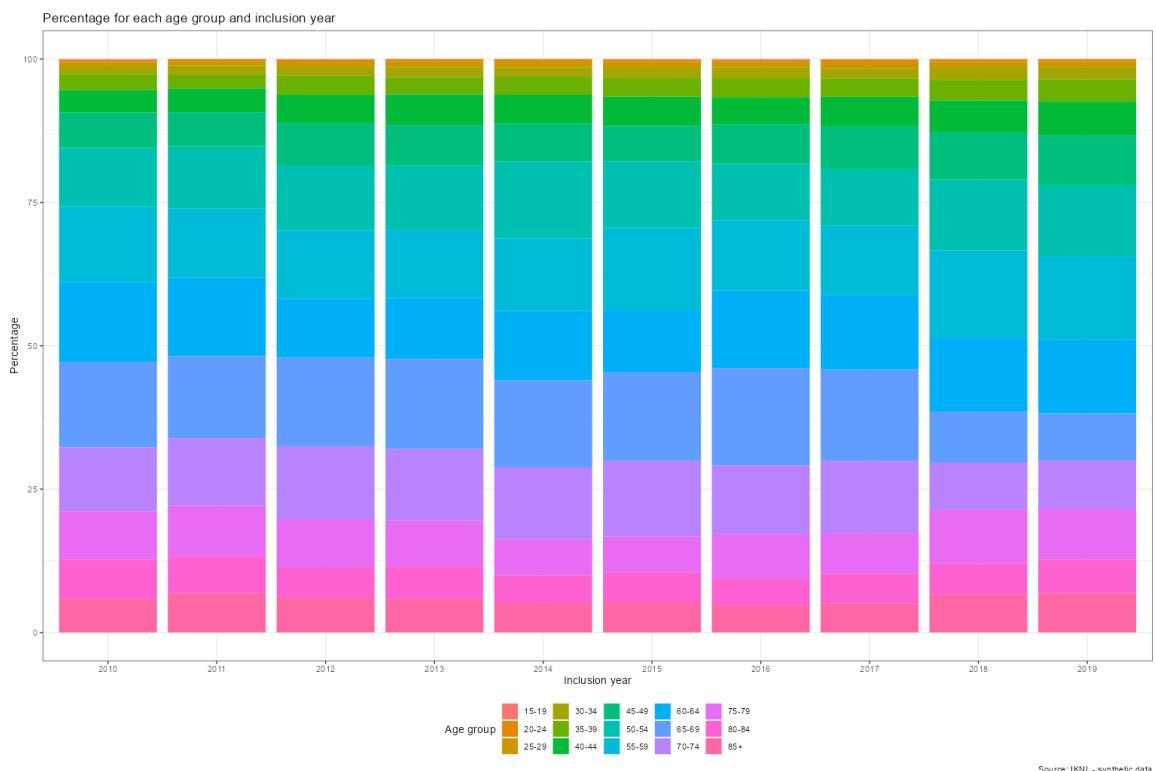


Figure 142. Distribution of age per inclusion year.

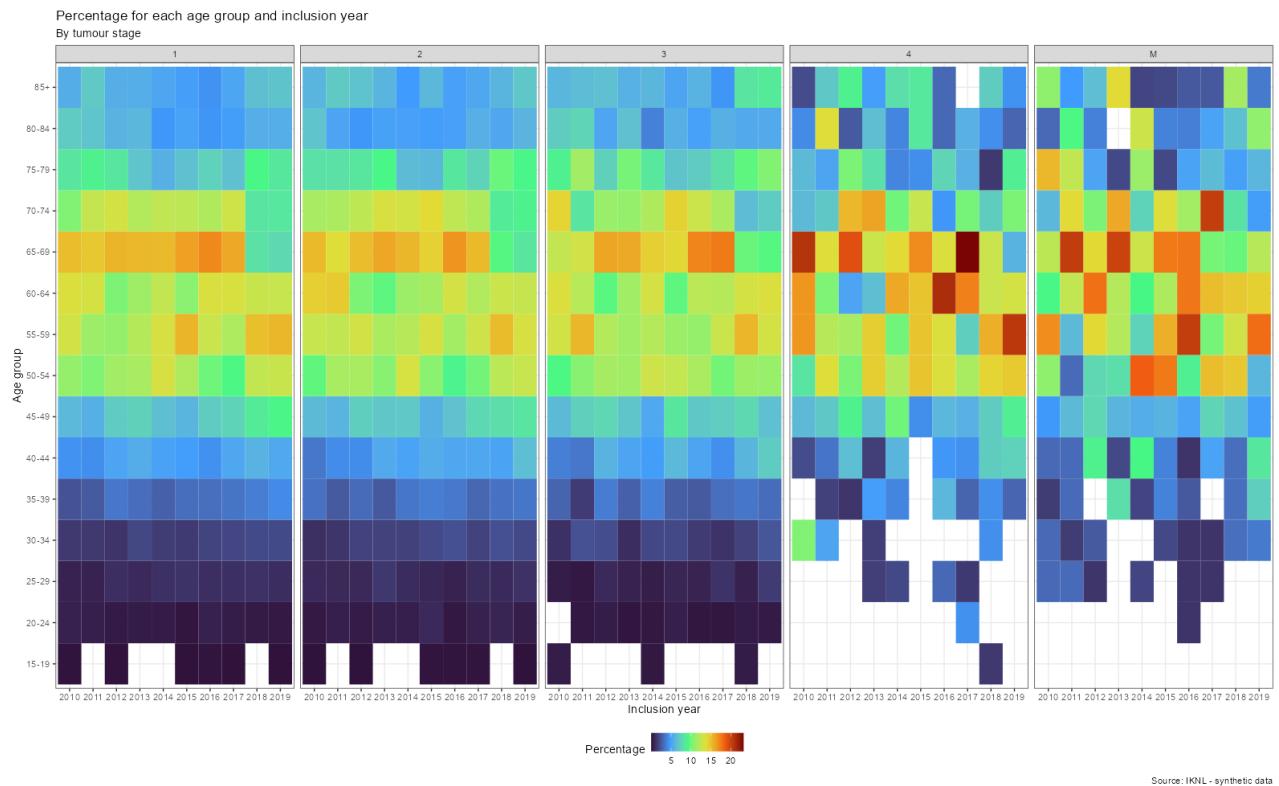


Figure 143. Distribution of age group per tumor stage and inclusion year.

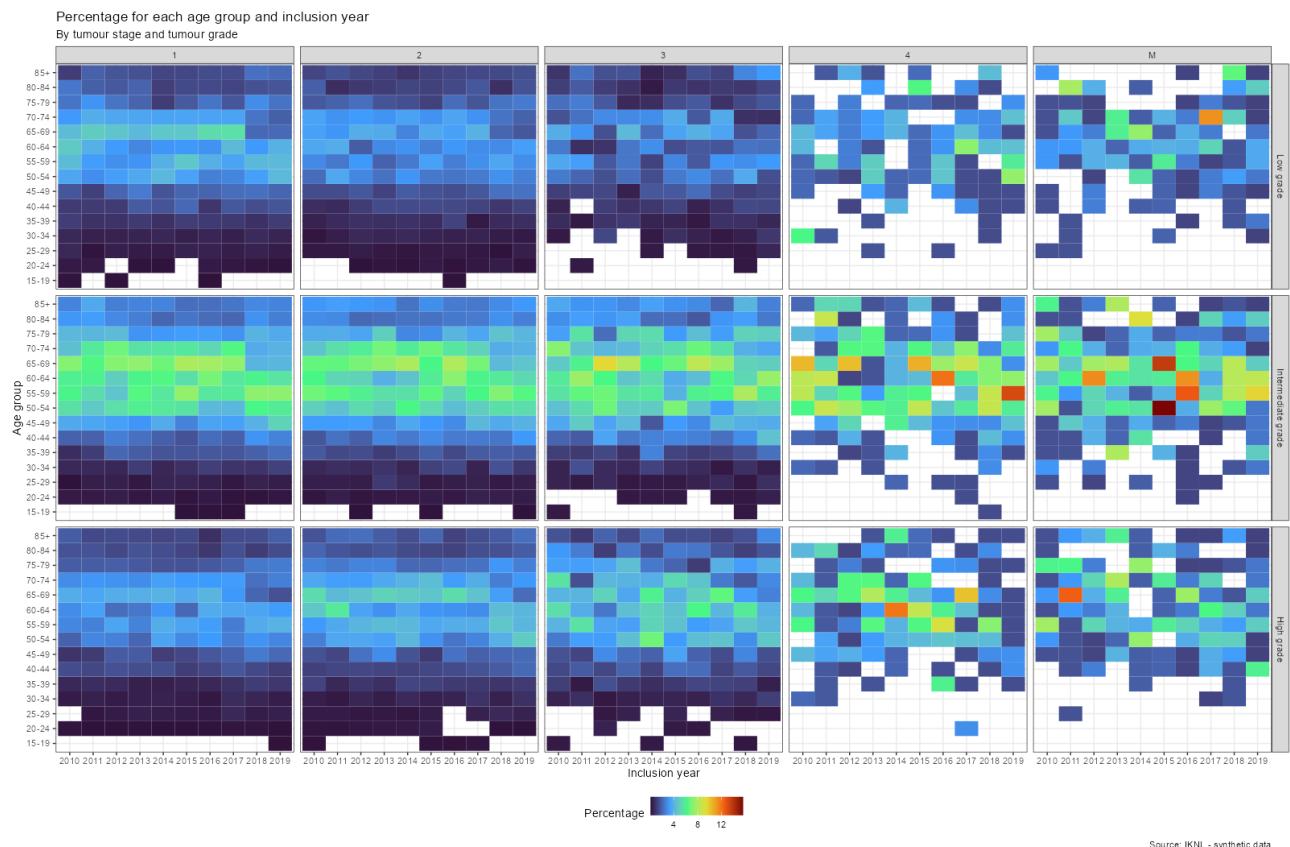


Figure 144. Distribution of age group per tumor stage, tumor grade, and inclusion year.

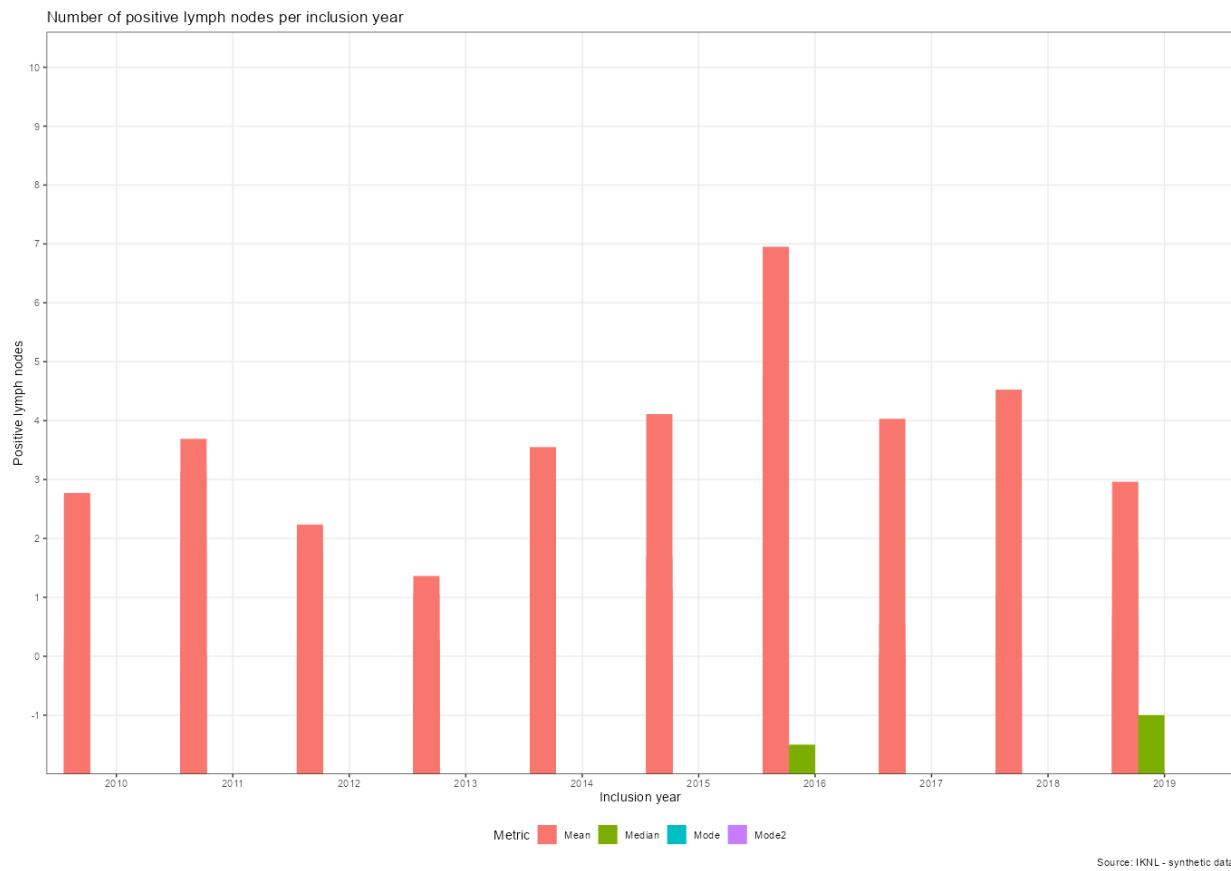


Figure 145. Number of positive lymph nodes per inclusion year using the mean, median and two different ways of measuring the mode.

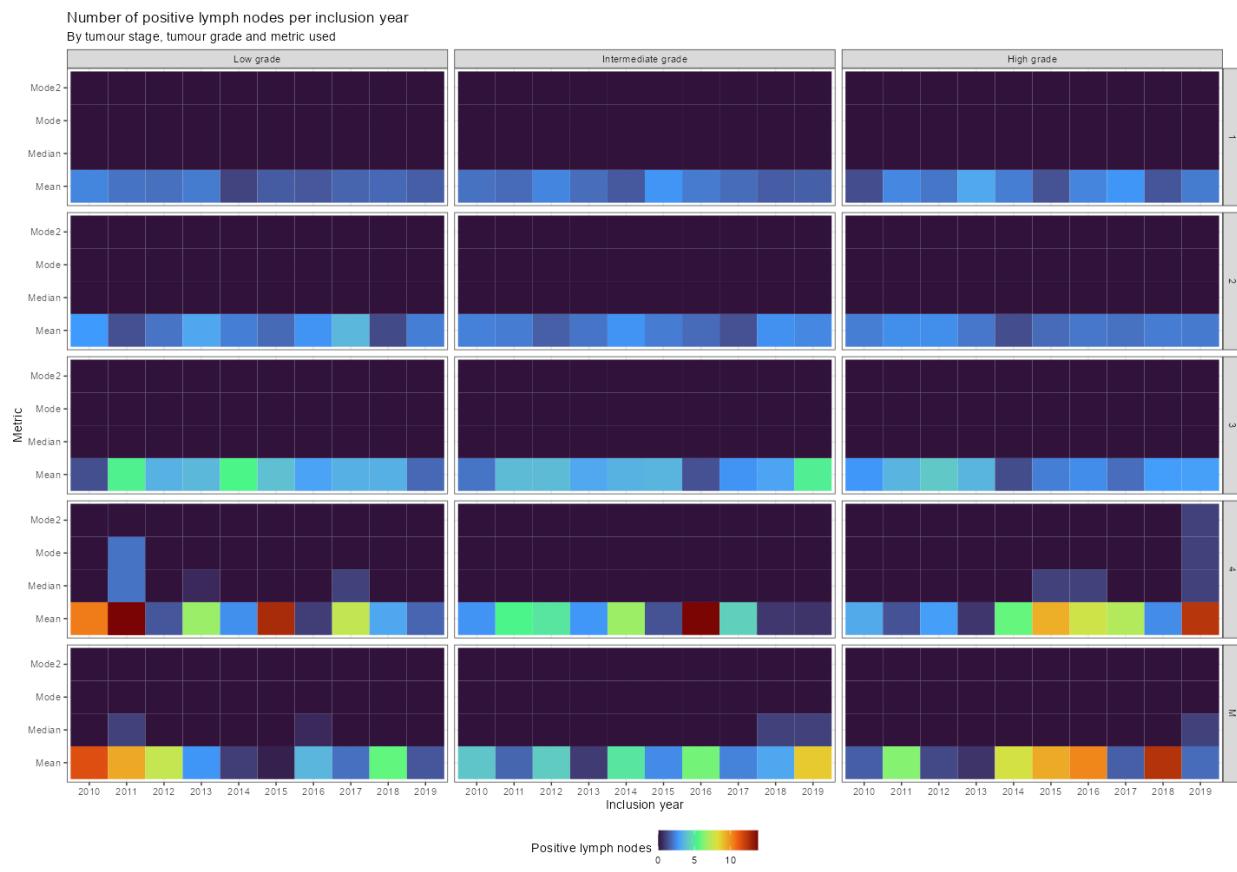


Figure 146. Number of positive lymph nodes per inclusion year using the mean, median and two different ways of calculating the mode.

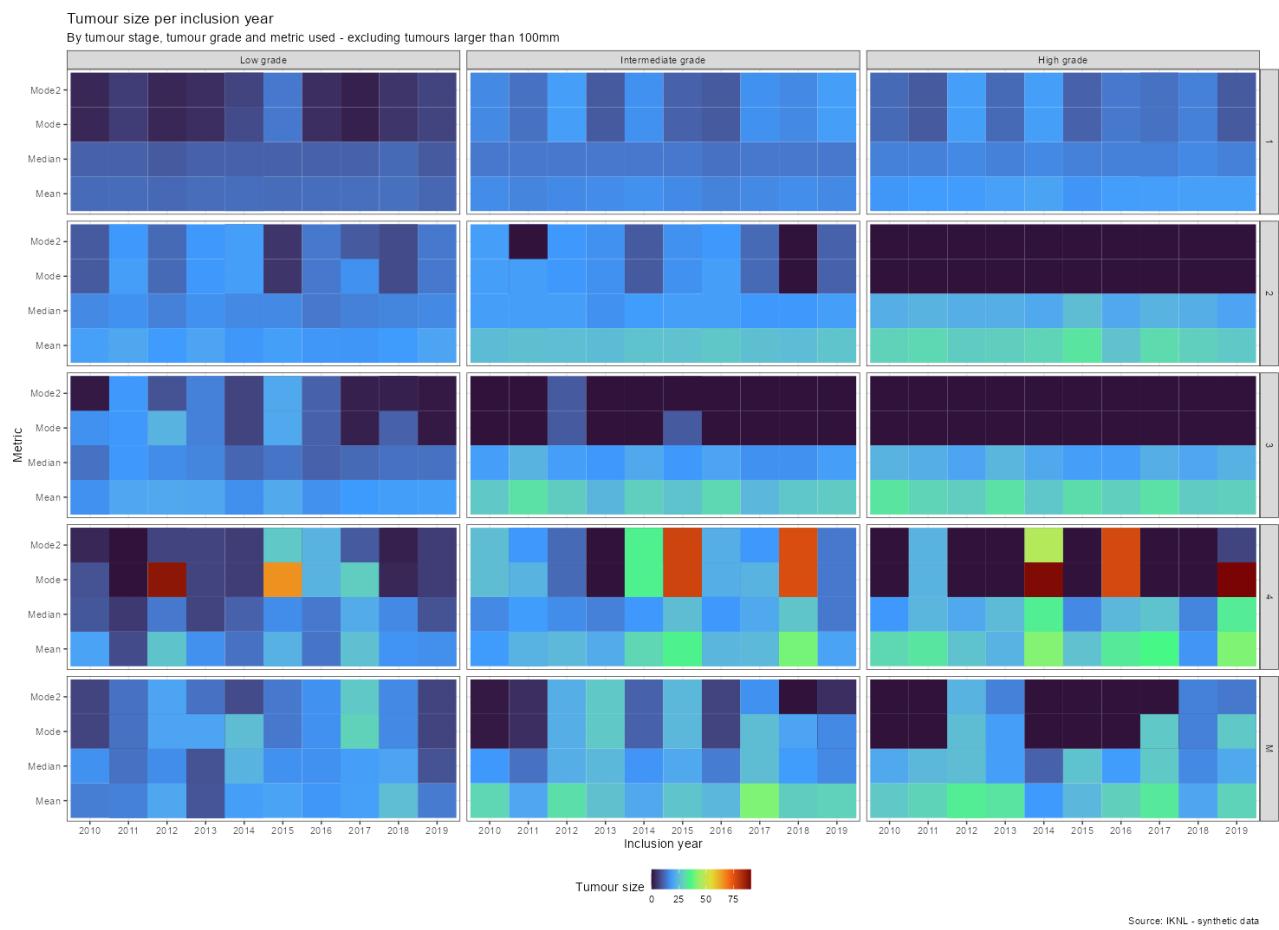


Figure 147. Tumor size per inclusion year by tumor stage, tumor grade and four different metrics used.

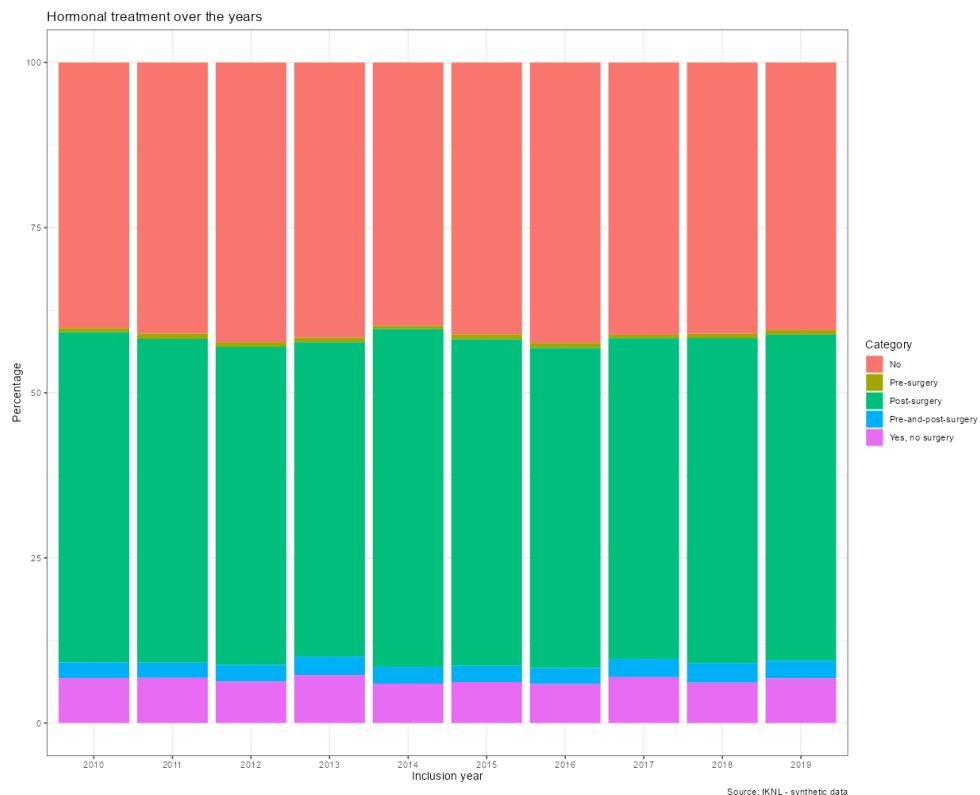


Figure 148. Distribution of hormonal treatment per inclusion year.

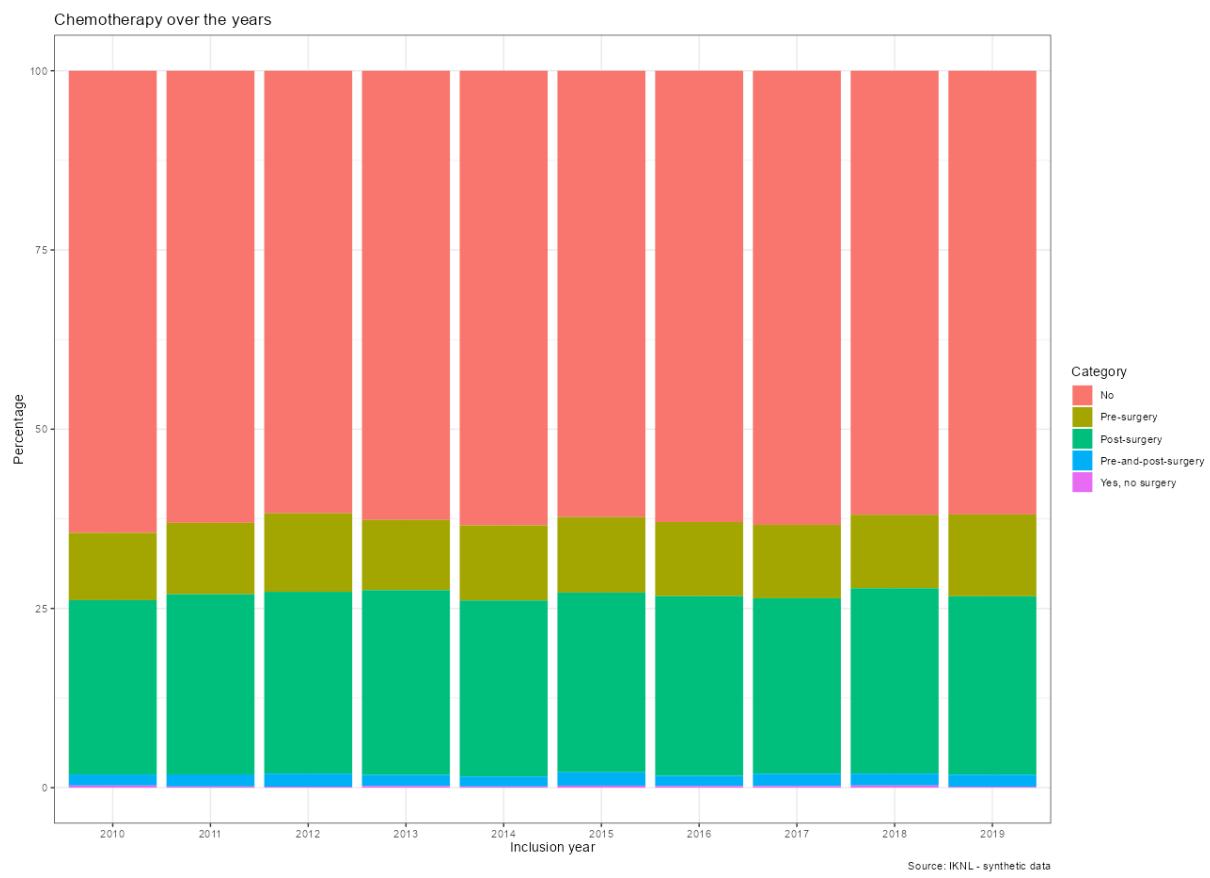


Figure 149. Distribution of chemotherapy per inclusion year.

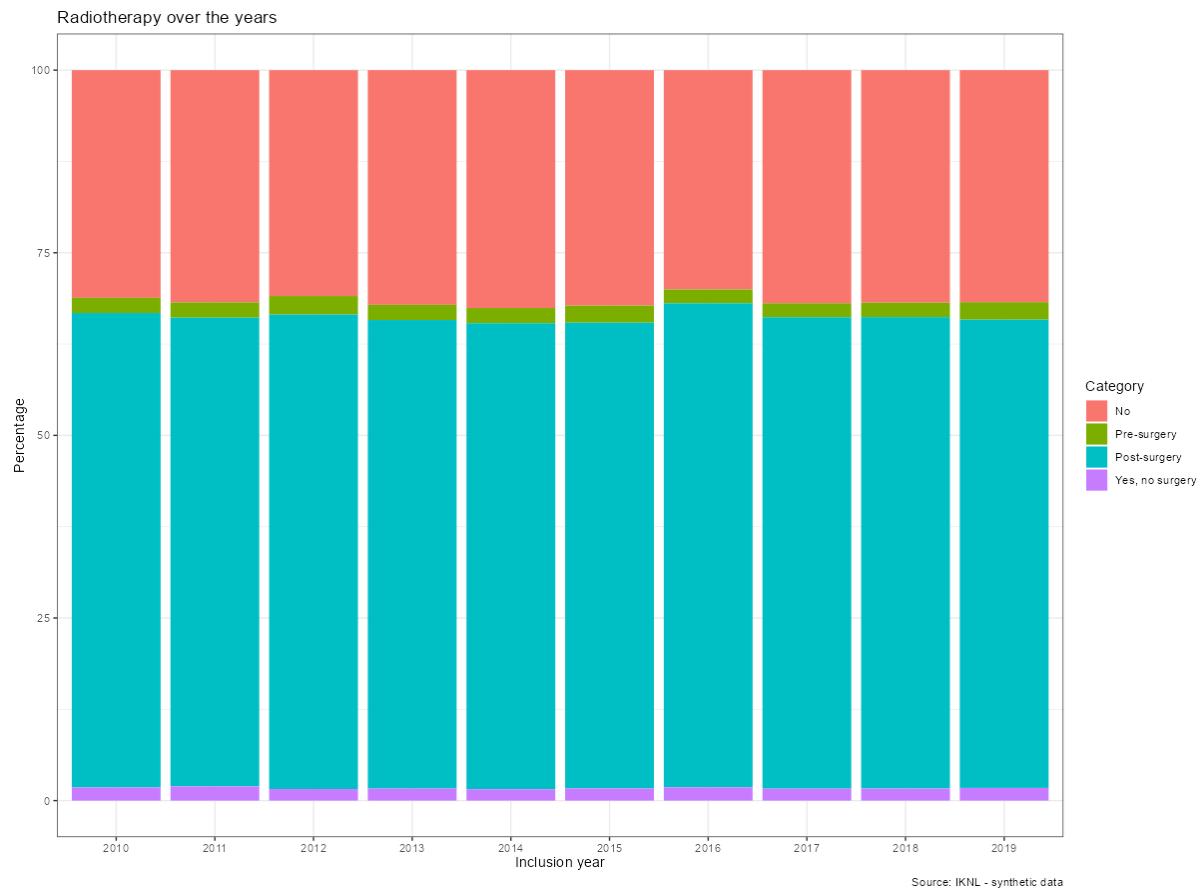


Figure 150. Distribution of radiotherapy per inclusion year.

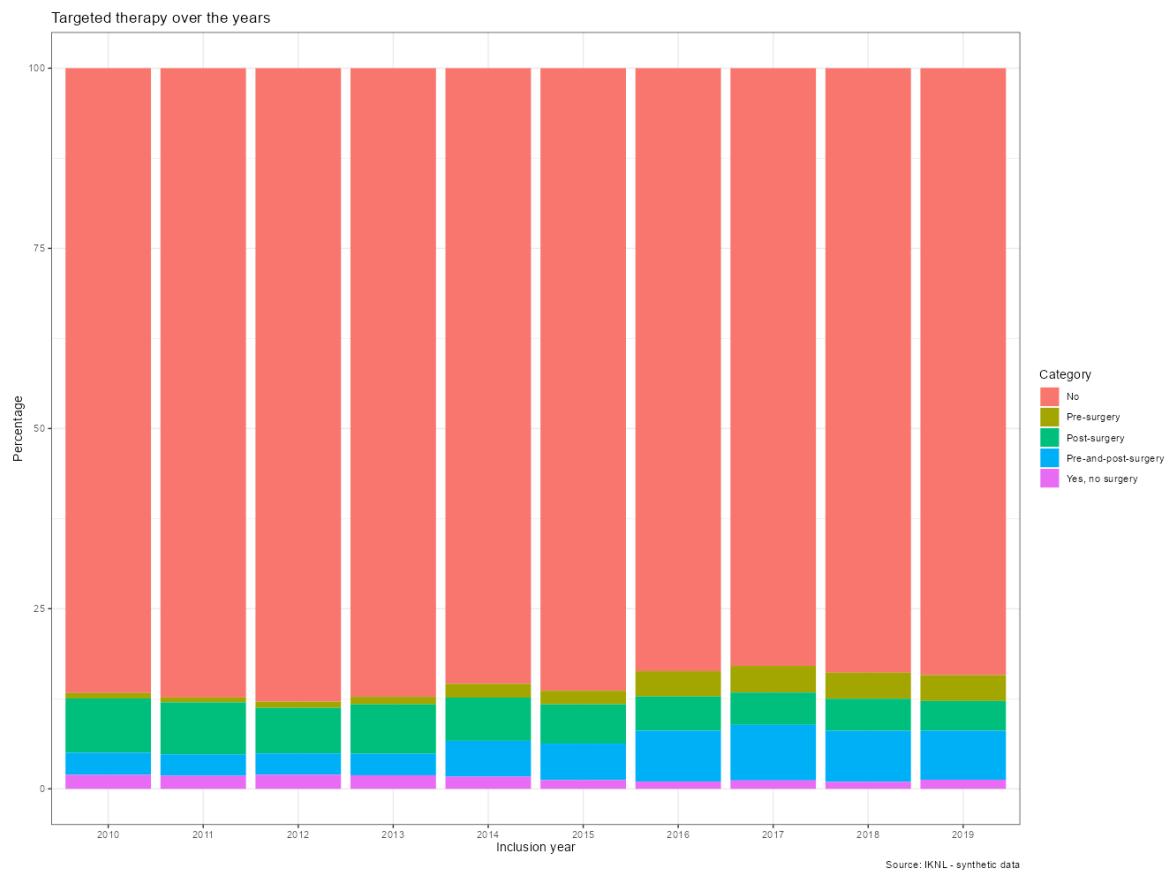


Figure 151. Distribution of targeted therapy per inclusion year.

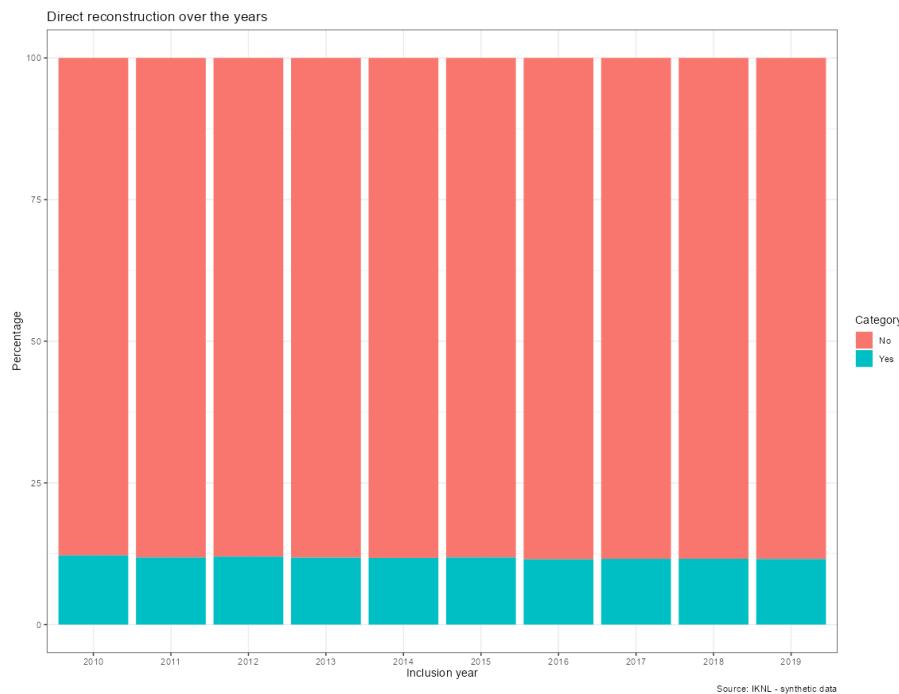


Figure 152. Distribution of direct reconstruction per inclusion year.

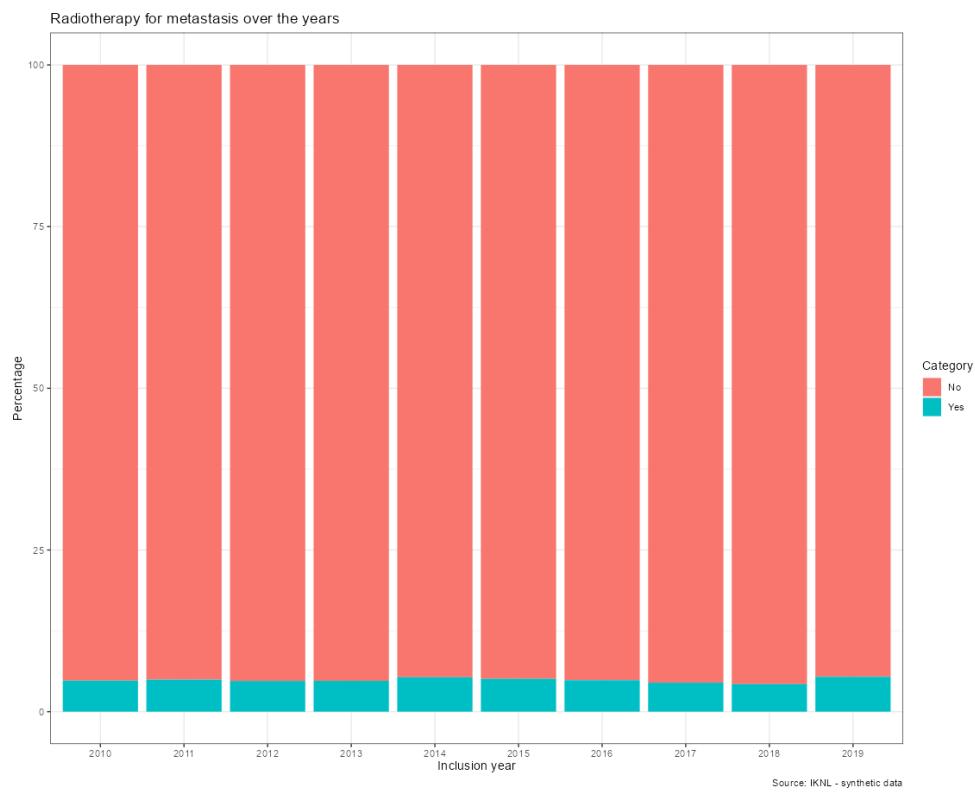


Figure 153. Distribution of radiotherapy for metastasis per inclusion year.

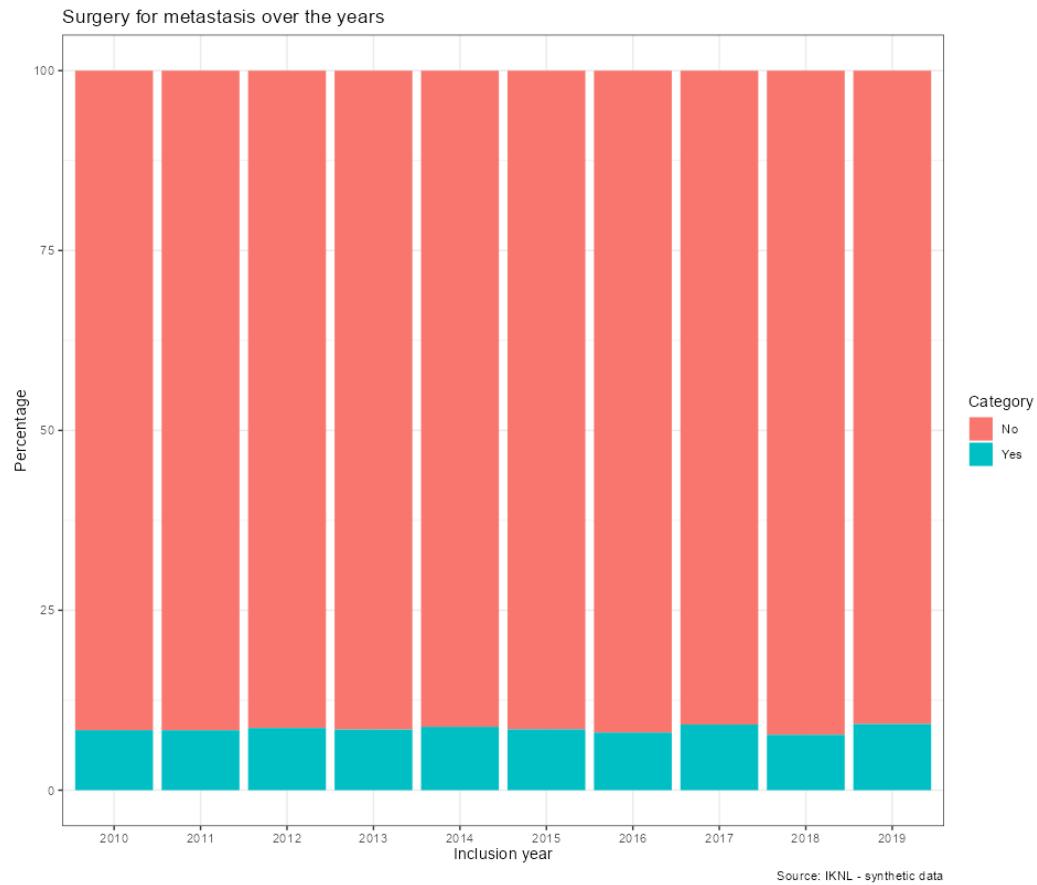


Figure 154. Distribution of surgery for metastasis per inclusion year.

Figure 155 shows a somewhat complex figure looking at how many patients receive combination treatment, which consists of possible hormonal therapy, chemotherapy, radiotherapy, and targeted therapy. We did not include surgery. It seems that the majority include radiotherapy and hormonal therapy post-surgery, or separately.

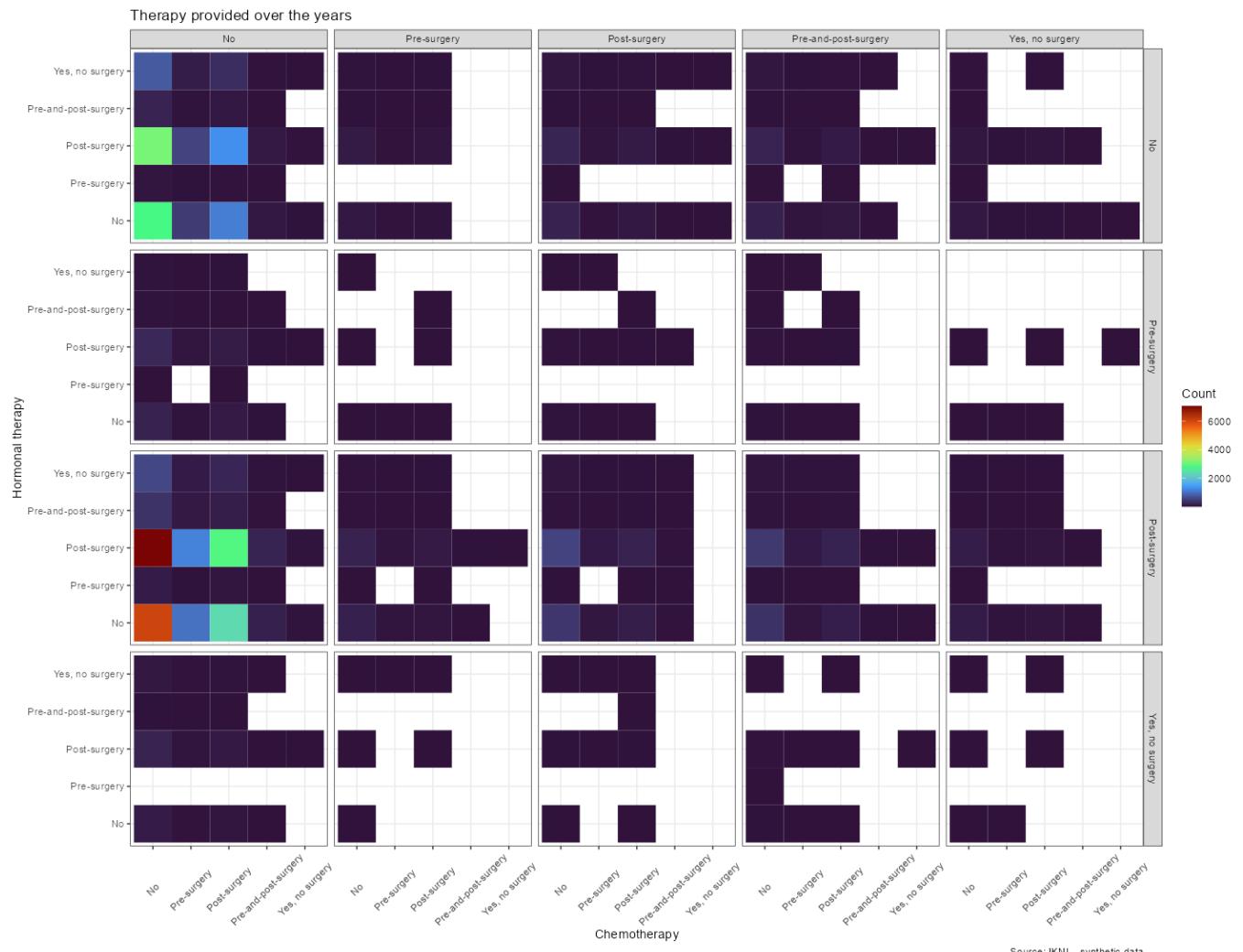


Figure 155. Distribution of therapy provided per hormonal treatment, chemotherapy, radiotherapy (facet rows), and targeted therapy (facet columns).

In addition, we want to use this data to look at what the potential ‘average’ breast cancer patient would be so we can use these characteristics for predictive modelling. Although there are numerous limitations to this approach, it does make sense to see which clinical parts are important, and if we see shifts across the past ten years (2010 – 2019). If we look at all the information of the synthetic dataset and combine this with the information necessary to run a developed prediction model.

In summary, it seems that the ‘average’ breast cancer patient is around 59 years of age, has a T2N1M0 15mm tumor of intermediate grade, with 3 positive lymph nodes and a progesterone and estrogen positive receptor. For treatment, many received surgery and

post-surgical radiotherapy and/or hormonal treatment. The majority did not receive chemotherapy, nor targeted therapy.

Clinical prediction models found.

In our rapid review, we identified CancerMath and PREDICT as validated prediction models, and therefore used both tools to look at the predicted survival rate for specific clinical groups. These groups are partly based on our efforts to show (and understand) how changing various predictors would change the predicted overall survival, and to show the predicted overall survival for the ‘average’ breast cancer patient (i.e., patients aged 59).

Estimating overall five-year survival using the NHS PREDICT tool ¹¹⁸							
DCIS or LCIS only?	No	No	No	No	No	No	No
Age at diagnosis	45	45	45	45	59	59	59
Post-menopausal	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
ER status	Positive	Positive	Positive	Positive	Positive	Positive	Positive
HER2/ERRB2 status	Positive	Positive	Positive	Positive	Negative	Negative	Negative
Ki-67 status	Positive	Positive	Positive	Positive	Positive	Positive	Negative
Hormone therapy	5 years	5 years					
Chemotherapy	3 rd gen	No					
Trastuzumab	Yes	Yes	Yes	Yes	N/A	N/A	N/A
Bisphosphonates	No	No	No	No	No	No	No
Invasive tumour size (mm)	5	5	10	10	15	15	15
Tumour grade	1	1	2	2	2	2	2
Detected by	Unknown	Screening	Unknown	Screening	Unknown	Screening	Screening
Positive nodes	0	0	5	5	3	3	3
Micro metastases	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Overall survival							
Surgery only	98%	98%	88%	90%	85%	87%	91%
+ Hormone therapy	99%	99%	91%	92%	89%	90%	93%
+ Chemotherapy	99%	99%	94%	95%	92%	93%	N/A
+ Trastuzumab	99%	99%	95%	96%	N/A	N/A	N/A
+ Bisphosphonates	99%	99%	96%	97%	93%	94%	N/A

Table 1. Results from estimating the five-year survival probability via the NHS PREDICT tool.

Estimating overall 15-year survival using CancerMath ¹¹⁹							
Current Age	45	45	45	45	59	59	59
Tumour diameter (cm)	0.5	1	1.5	2	1.5	1.5	1.5
Positive nodes	0	0	5	5	3	5	5
ER status	Positive						

¹¹⁸ <https://breast.predict.nhs.uk/tool>

¹¹⁹ <http://www.lifemath.net/cancer/breastcancer/therapy/index.php>

PR status	Positive	Positive	Positive	Positive	Positive	Positive	Positive
HER2 status	Positive	Positive	Positive	Positive	Negative	Negative	Positive
Histological type	Ductal	Ductal	Ductal	Ductal	Ductal	Ductal	Ductal
Tumour grade	1	2	2	2	2	2	2
Hormonal therapy	Tamoxifen	Tamoxifen	Tamoxifen	Tamoxifen	Tamoxifen	Tamoxifen	Tam. To AI
Chemotherapy	Dd CA*4 + T*4	none					
Cancer Mortality	0.9%	3.5%	12.9%	13.9%	10.8%	14.2%	23%

Table 2. Results from estimating the fifteen-year survival probability via the CancerMath tool.

Both CancerMath and PREDICT show a good survival rate for specific cancer patients that we ‘devised’ (Tables 2 and 3). Of course, both results are somewhat difficult to compare since PREDICT offers 5-year survival rates, whilst CancerMath looks at 15—year survival rates. It is commonly known, and we can see from the NKR data, that survival rates do not drop at a fixed slope. Hence, we cannot just backtrack the 15-year survival rate to the five-year survival rate. Nevertheless, both prediction tools show that if a T2N1M0 were diagnosed, and received their treatment, their survival rate would most likely extend beyond the first two years following diagnosis.

Comparison of clinical prediction founds and the clinical prediction model built Despite our strong objections for using the clinical prediction model we built, we still tried to compare it with the other models just to have a glimpse of how far off we would be (Table 3). In summary, it seems that regardless of model used, a patient with a T2N1M0 has a good five-year survival. We could even speculate that a patient seen by screening is a T1 with no positive lymph nodes and we would expect such a person to fair even better.

Perhaps the only way to find excess mortality explained via delayed breast cancer screening is by placing the 87% survival probability of the NHS PREDICT tool at the beginning of the curve. However, this would not make sense, especially if we look at the IKNL data¹²⁰, meaning that even if certain patients do not have a 95% five-year survival, there survival rate is often still quite high and especially so in the first two years following surgery.

Model used	NHS PREDICT tool	CancerMath	Our prediction model
Survival probability at	5 years	15 years	5 years
DCIS or LCIS only?	No	No	No
Age at diagnosis	59	59	59
Post-menopausal	Yes	N/A	N/A
ER status	Positive	Positive	Positive
HER2/ERRB2 status	Negative	Negative	Negative

¹²⁰ <https://iknl.nl/kankersoorten/borstkanker/registratie/overleving>

Ki-67 status	Unknown	N/A	N/A
Progesterone status	N/A	Positive	Positive
Histological type	N/A	Ductal	N/A
Hormone therapy	5 years	Tamoxifen	Post-surgery
Radiotherapy	N/A	N/A	Post-surgery
Chemotherapy	none	Dd CA*4 + T*4	Pre-surgery
Surgery	N/A	N/A	Lumpectomie (with OKR)
Trastuzumab	No	N/A	N/A
Bisphosphonates	No	N/A	No
Invasive tumour size (mm)	15	15	15
Tumour grade	2	2	2
Detected by	Unknown	N/A	N/A
Positive nodes	3	3	3
Micro metastases	N/A	N/A	N/A
Survival probability			
Surgery only	87%		93% (92% – 94%)
+ Hormone therapy	90%	88.5%	93% (92% – 95%)
+ Chemotherapy			94% (93% – 95%)
+ Trastuzumab			
+ Bisphosphonates			
+ Radiotherapy			94% (93% – 95%)

Table 3. Results from estimating the survival of a particular patient using three different models.

Literature research on the influence of delayed care during Covid-19 on breast, cervical and colon cancer mortality.

[Our rapid review on the influence of delayed care during Covid-19](#) on breast, cervical and colon cancer mortality shows that it is highly unlikely that the excess mortality in 2020 and 2021 can be explained by delayed care. Almost all papers read and reviewed agree that delayed care had a major impact on screening and the number of diagnoses. However, for cancer patients that were already diagnosed with cancer and treated for their cancer, the delay in care resulted in a shift. We did find some papers that describe a difference in their treatment, but none of the papers hinted at direct preventable death due to a shift in treatment or a delay in treatment, although many also stated that the future might show different figures. Our search did not focus only on papers looking at 2020 and 2021, but we included all papers up to 2023 that would provide us with insight on the effect of delayed care on.

Clinical expert opinion.

From each of the three clinical experts we received written statements based on the document we sent and the scenarios we hypothesized. Each document is copy-pasted below and only changed by us if the numbering of the figures did not align anymore.

Jan Bonte (neurologist)

If I understand the question correctly, I am asked how likely or unlikely I find each of the four hypotheses. I will therefore do so point by point.

First of all, it should be clear that determination that Covid-19 is the cause of death is not a one-dimensional entity. There are many degrees to which Covid-19 can contribute to the death of people, ranging from a minimal contribution to the root cause of death. However, when completing the death certificate, a primary cause of death must be given and then it must be indicated what the 'chain of events' was that led to the demise. And it is known from the literature that the misclassification rate can be as high as 35-40%.¹²¹ This is a persistent and recurring problem in recording the cause of death. And so that also proved to be a major problem in determining death caused by Covid-19.¹²² It therefore means that any registration of Covid-19 as a cause of death should be viewed with skepticism, especially at the start of the pandemic in the Netherlands there were an insufficient number of tests to confirm the diagnosis in nursing homes and in the home setting, and the mere suspicion of Covid-19 was enough to enter it on the death certificate, as home doctors and GPs were instructed.

1. There is no excess mortality in 2020 and 2021.

This hypothesis can be rejected on the basis of the aggregate data made available to me. There is excess mortality at multiple points in time and over an extended period (Figure 13), and the cumulative excess mortality over the 2020/2021 period is also elevated (Figure 16).

2. The excess mortality from 2020 and 2021 can be explained by Covid-19.

This hypothesis seems very plausible for the first peak in excess mortality at the time of the pandemic outbreak in the Netherlands in spring 2020. The broad peak in autumn 2020 and winter 2020/2021 is also very likely due to Covid-19. For these three peaks in excess mortality, it runs parallel to the mortality caused by Covid-19. This is also true for the peak in excess mortality in autumn and early winter in late 2021. This may be called extraordinary though, and I therefore wonder whether the excess mortality in that period can be explained entirely by Covid-19. Because in the meantime, a very large part of the population has been vaccinated against Covid-19, and certainly the vulnerable people, those with underlying diseases and the elderly. Among them, vaccination coverage was highest. If this spike in excess mortality were indeed caused entirely by Covid-19, it may be argued that the primary promise of the mass vaccination campaign has not been fulfilled: preventing mortality. This even more so because one would expect that in the first two waves the most

¹²¹ Smith Sehdev AE, Hutchins GM. Problems with proper completion and accuracy of the cause-of-death statement. *Arch Intern Med.* 2001;161(2):277-284. doi:10.1001/archinte.161.2.277

¹²² Armstrong D. The COVID-19 pandemic and cause of death. *Soc Health Illn.* 2021;43(7):1614-1626. doi:10.1111/1467-9566.13347

vulnerable people had already died, and one would not expect another high peak in excess mortality to follow.

There is another peculiar thing to see when looking at the graphs that break down excess mortality by age. It is important to note that the y-axes of the three graphs differ. If one takes that into account, it is clear that for the first one those peaks in excess mortality are highest among the over-80s, followed by the 65-80 age group. The excess mortality is lowest in the age group of people younger than 65 years. There is some excess mortality in that age group in autumn 2020 and winter 2020/2021, but it is minimal. That will change in autumn 2021. Suddenly, there is a spike in excess mortality in this age group that is even higher and wider than at the time of the outbreak. This is completely contrary to expectations.

Meanwhile, a significant proportion of this age group had gone through the infection - resulting in excellent protection against reinfection¹²³ - and a significant proportion of them had also been vaccinated against Covid-19. It is extremely curious that, taking these two facts into account, that the excess mortality in this group would be higher than in previous waves. In my opinion, this is in no way consistent with the claim that the over-mortality in this group would be due to Covid-19. A better explanation will therefore have to come here.

3. The excess mortality from 2020 and 2021 can partly be explained by the temporary stop of population screening.

In my opinion, this hypothesis can be safely rejected. Screening for breast, colon and cervical cancer over decades has never been able to demonstrate a convincing decrease in all-cause mortality.¹²⁴ A recent meta-analysis reaffirms this.¹²⁵ If screening programs have not shown a reduction in mortality over decades, it would be very extraordinary if missing only one screening moment suddenly led to this significant excess mortality. The screening intervals for breast, cervical and colon cancer are two to three years, five years and two years, respectively. It is known that screening mainly leads to the detection of asymptomatic and relatively benign abnormalities with a much better prognosis (the length-time bias) than abnormalities found in the interval of two screening moments. Again, it would then be a miracle that missing a single screening moment would lead to increased mortality in such a short period of time. Even if a relatively benign tumor was missed because of one missed screening moment, it is very questionable that it would eventually lead to death, and is extremely unlikely that it would do so in short amount of time. Moreover, the vast majority of excess mortality concerns the age group over 80 years, at a time when the screening

¹²³ Gazit S, Shlezinger R, Perez G, et al. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Naturally Acquired Immunity versus Vaccine-induced Immunity, Reinfections versus Breakthrough Infections: A Retrospective Cohort Study. *Clin Infect Dis.* 2022;75(1):e545-e551. doi:10.1093/cid/ciac262

¹²⁴ Prasad V, Lenzer J, Newman D H. Why cancer screening has never been shown to “save lives”—and what we can do about it *BMJ* 2016; 352 :h6080 doi:10.1136/bmj.h6080

¹²⁵ Brethauer M, Wieszczy P, Løberg M, et al. Estimated Lifetime Gained With Cancer Screening Tests: A Meta-Analysis of Randomized Clinical Trials. *JAMA Intern Med.* Published online August 28, 2023. doi:10.1001/jamainternmed.2023.3798

program for cervical carcinoma has already ended 20 years, and the screening program for breast cancer has already ended five years. So those people have not missed a screening moment, because for them screening had already ended years ago. So, missing a screening moment could not have caused their deaths, because they simply didn't participate in the screening program anymore. Another argument is that screening for breast and cervical cancer by definition only affects women. However, excess mortality affects both men and women. How then to explain the excess mortality in men?

These arguments Figures 35, 40, and 45 convincingly summarizes my position. There is no increased mortality from the three types of cancers screened for in the Netherlands. It neatly follows the trend of previous years.

4. The excess mortality from 2020 and 2021 can partly be explained by delayed care of already diagnosed cancer patients.

It is somewhat more difficult to justify why even this hypothesis is most probably incorrect. First, doctors are generally well placed to know which people need urgent treatment after cancer diagnosis, and in which people some waiting time makes no difference to prognosis. Moreover, urgent cancer care has continued as much as possible, including surgery, radiotherapy, and chemotherapy. Again, even if a longer waiting time would lead to a higher cancer stage, it is highly unlikely to lead to such a substantial excess mortality in the short term. And again, Figures 35, 40, and 45 contradict this hypothesis. There is no increase in cancer mortality of the three main cancer types in the Netherlands, except lung cancer, which is not shown in the figure. And again, I cannot think of an explanation of how delaying cancer care would lead to waves in excess mortality. That goes against all laws in medicine. What could theoretically be is that a longer waiting time could lead to higher tumor stages and worse prognosis, and consequently higher mortality. But that is theory, because that mortality would only occur after a longer period of time and would not come in waves. And again, Figures 35, 40, and 45 provide no evidence for that hypothesis either.

Roderik Kraaijenhagen (cardiologist) and Sabine Pinedo (internist)

The following are the conclusion of the last slide of their presentation, which can be found [here](#):

1) Looking at the graphs we can conclude that this hypothesis is not true.

2) To answer this hypothesis, it is relevant to have some additional information, i.e.:

- What is the definition of 'caused by covid'? Is it caused by the disease itself or also due to for example:
 - the possible interaction of Covid with other diseases;
 - the lockdowns (and suicides) as result of the Covid-pandemic;

- undertreatment as result of misinterpretation of, for example, cardiac diseases because certain complaints are interpreted as complaints due to Covid,
 - vaccination for Covid due to the Covid-pandemic;
 - undertreatment because people were afraid to go to the hospital as they were scared to get Covid there.
- What was the procedure for concluding that a patient died of COVID?

Interestingly enough: in dec 2022-jan 2023 there is a peak in excess mortality, however there is no peak in Covid19 diagnoses. This is different as compared to the previous 2 years. As compared to 2020, at least 3 aspects were different in dec 2022-jan 2023; there was less testing for Covid in 2023, there was a booster campaign just before this period and there was no lockdown.

- 3) As explained in the presentation this hypothesis is not true. If there would be a measurable effect on excess mortality, the effect would be very small as compared to the excess mortality shown in the graphs.
- 4) This hypothesis is highly unlikely as the number of patients dying from cancer is not increased in 2020-2021 (see Figures 35, 40, and 50)

Discussion

The road from acceptance of this study (December 13, 2022) to delivering our report (October 13, 2023) was rocky. This has mainly to do with the current data landscape of The Netherlands which made it impossible to obtain the necessary patient-level data for answering our research questions. Although we believe that we would have been able to provide a much more detailed and methodologically sound answer had we been given the data we deemed necessary, we also felt that it was our duty to provide as much information as possible from the data freely available to us. However, this also meant that we had to diverge from our proposed vision to answer three research questions and instead focus on just on the one we deemed most important: *What is the effect of the Covid-19 pandemic on cancer screening in The Netherlands and how did this impact (excess) mortality in 2021?* Everything that we have done, and will discuss in the following pages, was aimed at trying to answer this research question. However, we added delayed care as well.

Considering the rocky road we experienced, we believe it is best to discuss our findings part-by-part starting with our experience with the data landscape in the Netherlands. We will then go on to discuss four possible scenarios that may contribute to the excess mortality in the Netherlands to 2020 and 2021.

Data landscape in the Netherlands

Health data is not data that is easily shared. Without a doubt it is very personal data showing for each patient their history in health care. For years, healthcare has been trying to connect more and more data sources enticing patients to share as much of their data as possible to healthcare providers and institutions. Not only it is easier for a doctor to treat a patient if the history of that patient is readily available, but it may even save lives. For institutions, and companies alike, that data is where a tremendous amount in scientific knowledge and market share enables the creation of tailor-made programs and interventions (i.e., psychological, pharmaceutical, surgical etc.). Last, but not least, health insurance companies would greatly benefit when trying to determine the optimal reimbursement price. Hence, one can easily see why such data is not openly available.

However, data that is not being openly available because it is sensitive is one thing, but data that can only be freely used by selected institutions for research and funding is another. When an institution has that amount of power yet still depends on outside funding it cannot be said that, without a doubt, this institution acts independently.

For cancer data, the Netherlands Comprehensive Cancer Organization (IKNL) is the only institution that has all the patient-level cancer data available. By connecting with other institutions, like CBS or RIVM, and by retrieving clinical data from almost every hospital in the Netherlands, the IKNL has an impressive database at the patient level from which it can draw many publications.¹²⁶ In addition, IKNL can easily connect with pathological data coming from PALGA.¹²⁷ These publications are of course extremely valuable: not only to advance science but also to feed information back to the healthcare system itself. As always, progress benefits from knowledge sharing.

From a scientific perspective, the most beneficial road for the public is by sharing data and codes across various parties. Information dissemination via publication alone is severely limited as publications only tend to report positive findings (i.e., publication bias). Word limitations also make it more difficult to precisely read (or describe) what has been done. Hence, data and code sharing are the scientific way to go. It is here that IKNL advances cancer care by creating and easily offering synthetic data. Despite the inherent limitations of synthetic data, it still offers a fruitful way of sharing methods and insights.

For our research question, however, we needed more data and for different cancer types. It is here that data sharing no longer acted as the most beneficial road for an individual institution or university. We tried to obtain data from CBS, and we tried to obtain data from

¹²⁶ https://iknl.nl/getmedia/8e59581f-f175-4cce-88ec-8dacdd215ed3/2020_IKNL_Overzicht-wetenschappelijke-publicaties.pdf

¹²⁷ <https://www.palga.nl/>

IKNL. Both efforts failed, but we deemed the latter effort to be most detrimental. This is because the IKNL datasets are almost tailor made to answer our research objectives. Unfortunately, our request was denied multiple times. Officially, because our research objective was not clear, and we did not have the right type of clinicians involved.

We suspect, however, that the true reason for declining our request was a serious conflict of interest. In the [introduction section](#) we already mentioned that at the time of our request IKNL was writing a [research proposal](#) of their own which was very closely related to ours. Had we been given the data we would have published the results sooner, making use of the exact same database.

Despite popular belief, science is not always about collecting information, but about presenting (any) findings that are novel and exciting, first. And although science thrives by replication, this is seldom done because it is much harder to publish a paper that replicated another study. At the very least it is not as exciting. Thus, had we been given the data first and used that data to write this report, we would have ‘published’ first.

A denial of a request to share with IKNL our information, analytics, and findings to the highest degree possible (e.g., sharing statistical codes) shows that there was no intention to work together. We believe that working together would have benefited the public greatly, as research benefits from [both replication and enhancement](#).

It cannot be stated often enough that instead of receiving the data we deemed to be of the highest quality, and thus the most beneficial to the public, we now had to resort to ecological data freely available to all. It is quite telling that a country such as the Netherlands, which prides itself on research rigor and data availability, does not easily allow independent researchers to obtain the data they deem necessary - not even for a project granted by a Dutch governmental subsidiary body such as ZonMw. In fact, in this case, ZonMw may benefit from some reflection themselves. For researchers, funding means everything and by actively funding two similar projects, separately, ZonMw created tension which they should be aware of.

[Four possible scenarios](#)

Following the data we could access, and our initial research questions, we drafted four possible scenarios for which we could assess the level of evidence. These are:

1. There is no excess mortality in 2020 and 2021.
2. The excess mortality from 2020 and 2021 can be fully explained by Covid-19.
3. The excess mortality from 2020 and 2021 can be (partly) explained by the temporary cessation of population screening.

4. The excess mortality from 2020 and 2021 can be (partly) explained by delayed care of already diagnosed cancer patients.

For each of these questions, we assessed to what extent the data provided evidence in support of a hypothesis.

[There is no excess mortality in 2020 and 2021.](#)

This may seem a somewhat strange hypothesis to discuss first as the sole reason for the entire research line on excess mortality is based on the acknowledgement that there is excess mortality. However, it should be noted explicitly that excess mortality is not something which we observe but which we model (Figure 156). Hence, excess mortality numbers are just calculations based on subtracting the observed mortality from the expected mortality (Figure 157).

Excess mortality is not something that is easily defined or calculated. Although descriptions exist on which factors should be included, and it seems that the numbers calculated are done in a form of European collaboration, it is not clear how it is done exactly. What can be observed though, from plotting the data, is that the expected mortality forms a highly cyclical pattern with a slight positive slope. Hence, from the available data on expected mortality numbers, one can expect that seasonality (and thus infections) play a role as well as the slowly ageing population. There are no projections made separately for males of females, although it is widely known that females live longer. The entire goal of expected mortality calculations is to provide countries with the numbers of death they may expect, given a handful of variables, in a situation considered ‘business-as-usual’.

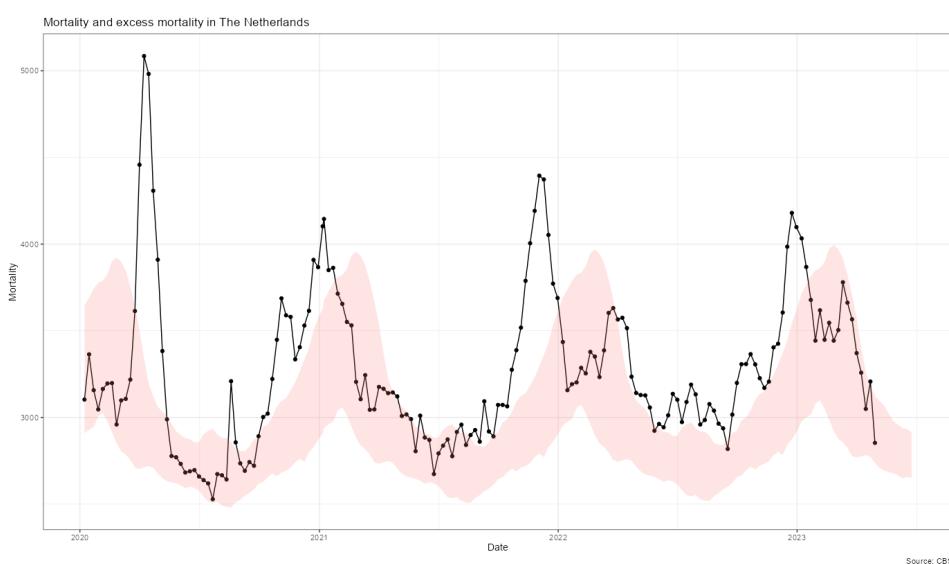


Figure 156. Mortality (black line) and expected mortality (red area) for the Netherlands. The excess mortality is the difference.

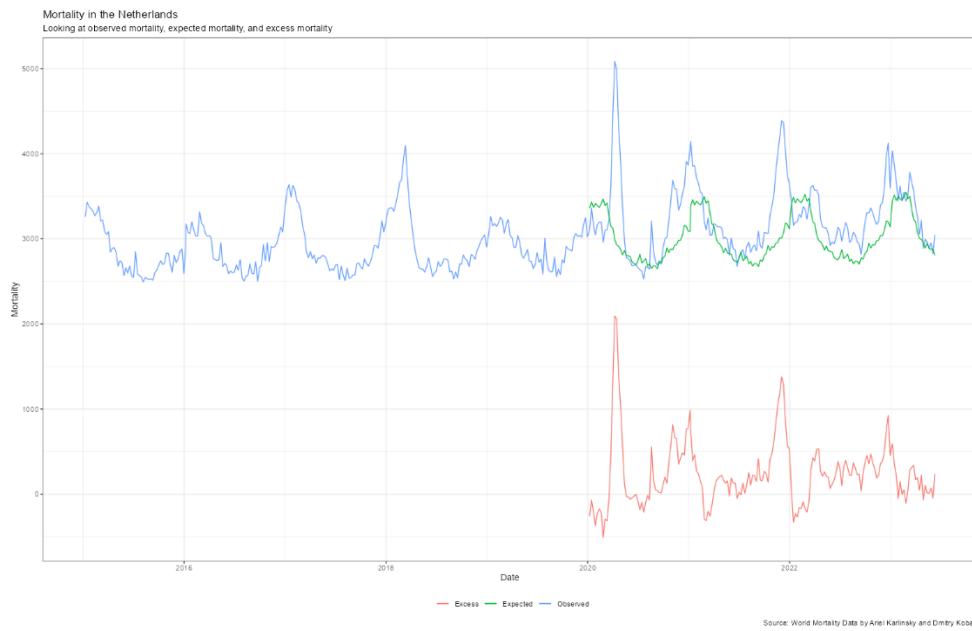


Figure 157. Excess mortality, expected mortality and observed mortality for the Netherlands.

Of course, Covid-19 was not ‘business-as-usual’ and the observed mortality spiked at the time Covid-19 was first detected in the Netherlands (March 2020). Therefore, the excess mortality numbers make sense in the earliest phase of the pandemic. However, Figure 156 also shows why the current expected mortality rate is outdated. First, the expected mortality does not really change across the years and certainly did not change due to Covid-19. Although the CBS life expectancy model did change (Dutch citizens born from 2020 onwards have a lower expectancy (according to the CBS) than their peers born a couple of years ago; see Figure 158), the expected mortality models did not. This is why there are so many periods in which the observed mortality exceeds the expected mortality.

Another peculiarity is the confidence bands surrounding the excess mortality metric: these are exceptionally stable. Most intervals of variation show stability as an inherent consequence of the model (the model is built on the complete dataset, and not continuously split and updated) so this is not something new, but we would expect more variation considering other infectious diseases such as the flu. For instance, 2018 also showed a severe mortality spike, second to the 2021 spike and third to the 2020 spike (Figure 157).

Also, one can clearly see from the data that mortality has been steadily rising for years as a function of age. Keeping the expected mortality almost the same would inevitably lead to an increase in excess mortality.

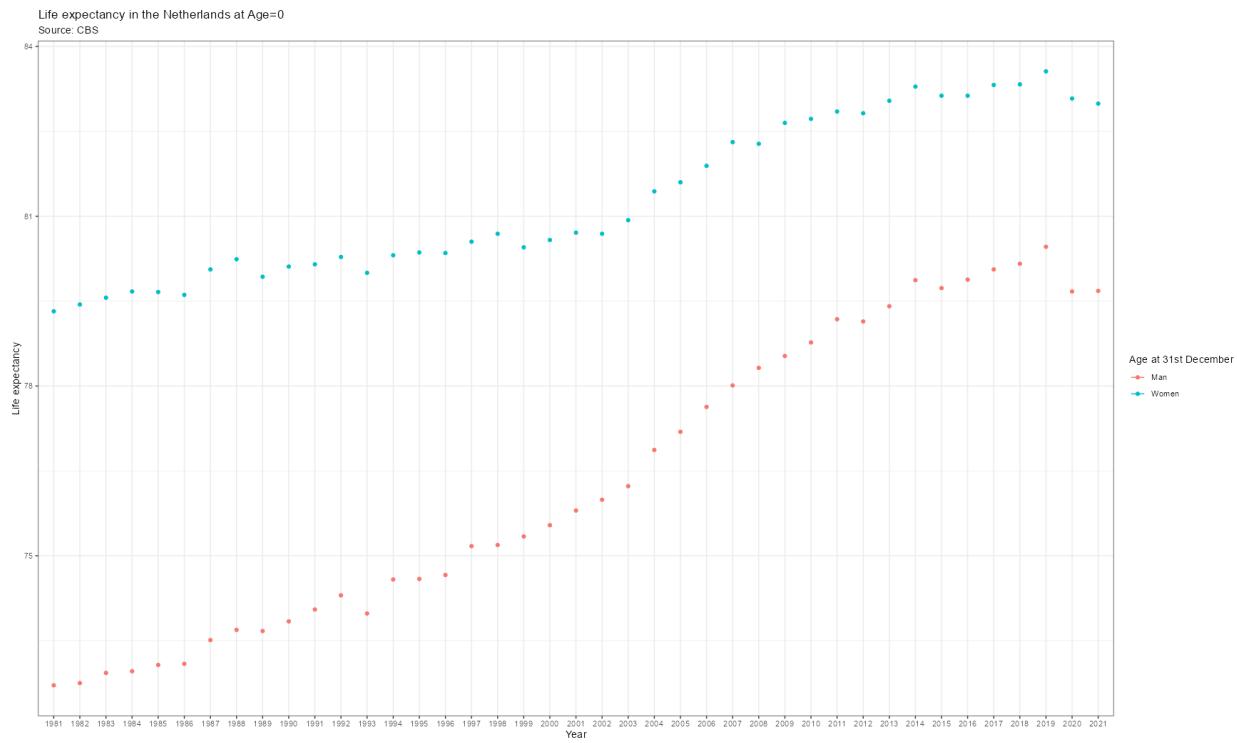


Figure 158. Life expectancy of someone age zero at the 31st of December per gender and year (1981 – 2021).

By far the greatest limitation of the excess mortality metric is the model itself. Several models on expected mortality exist which, automatically, lead to different numbers (the observed mortality never changes, only the expected mortality can change). Would a model automatically include Covid-19 expected deaths, it is very likely that the excess mortality metric would be much lower in 2021. One could even argue that including the deaths of 2020 in a model used to estimate the expected deaths of 2021 would cancel out any excess mortality. However, such an exercise would cancel out the entire aim of the expected mortality metric: to present a ‘business-as-usual’ scenario for identifying out-of-place scenarios.

Does this mean that there is no actual excess mortality? We believe that it does not. Especially for 2020 it is clear to see that we have much higher mortality than before. Also 2021 shows a mortality spike which is most likely larger than may be expected. We therefore do not believe (but cannot fully test since we do not have access to the excess mortality model) that any excess mortality is just bad modelling. Considering its widespread and consistent use, any inconsistency in mortality will immediately be picked up. We believe that this is one the major purposes of the excess mortality metric.

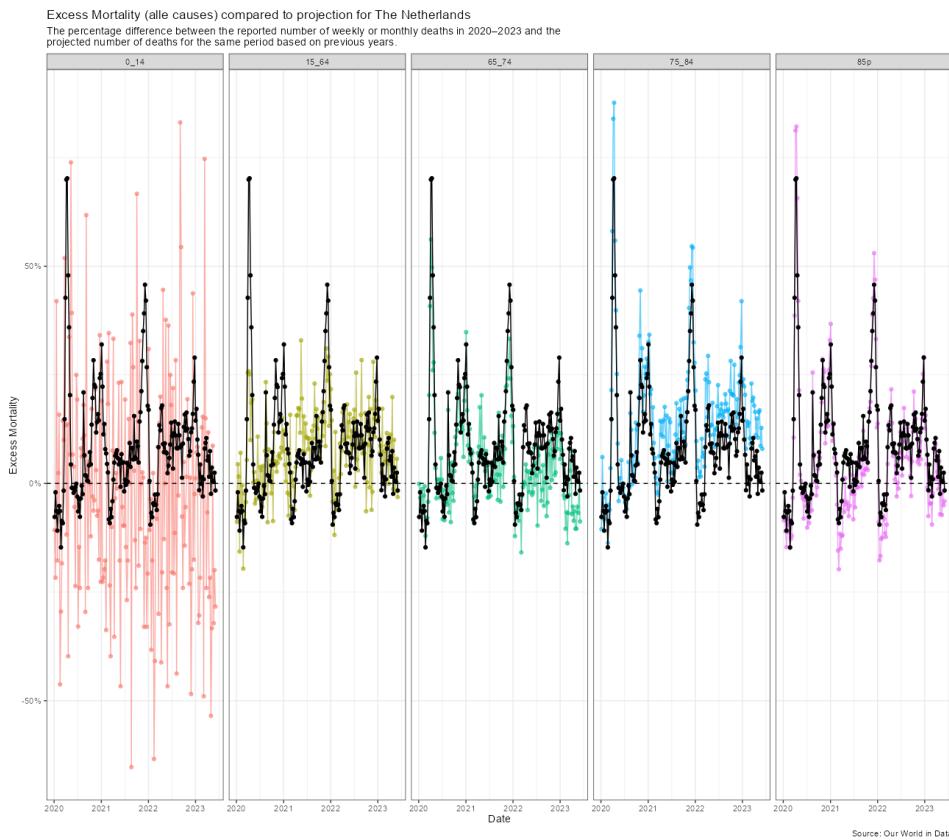


Figure 159. Excess mortality compared to projected mortality as the percentage difference between the reported number of weekly or monthly deaths and the projected number of deaths for the same period based on previous years per age category.

Even if this makes the expected 2020 values quite clear, it does nothing for the 2021 values. Had we included Covid-19 into the model we would have received different 2021 values for sure. Hence, the actual excess mortality number is not that important to us. We cannot just go and look for excess mortality numbers, look for the causes of death of the observed mortality and then see if numbers are missing or causes should be changed (Figure 160). The only thing we can do is see how the excess mortality numbers changed across time and find the best possible explanations coming from known biological pathways. Hence, for 2020 and 2021, we know that Covid-19 was present, cancer care was delayed (2020), social distance measures and masks mandated were introduced (2020 and 2021), and that vaccines were introduced (2021). These are by far the biggest changes in the years 2020 and 2021 and regardless of the correct excess mortality number, the bigger question is to what extent any of them contributed to observed mortality.

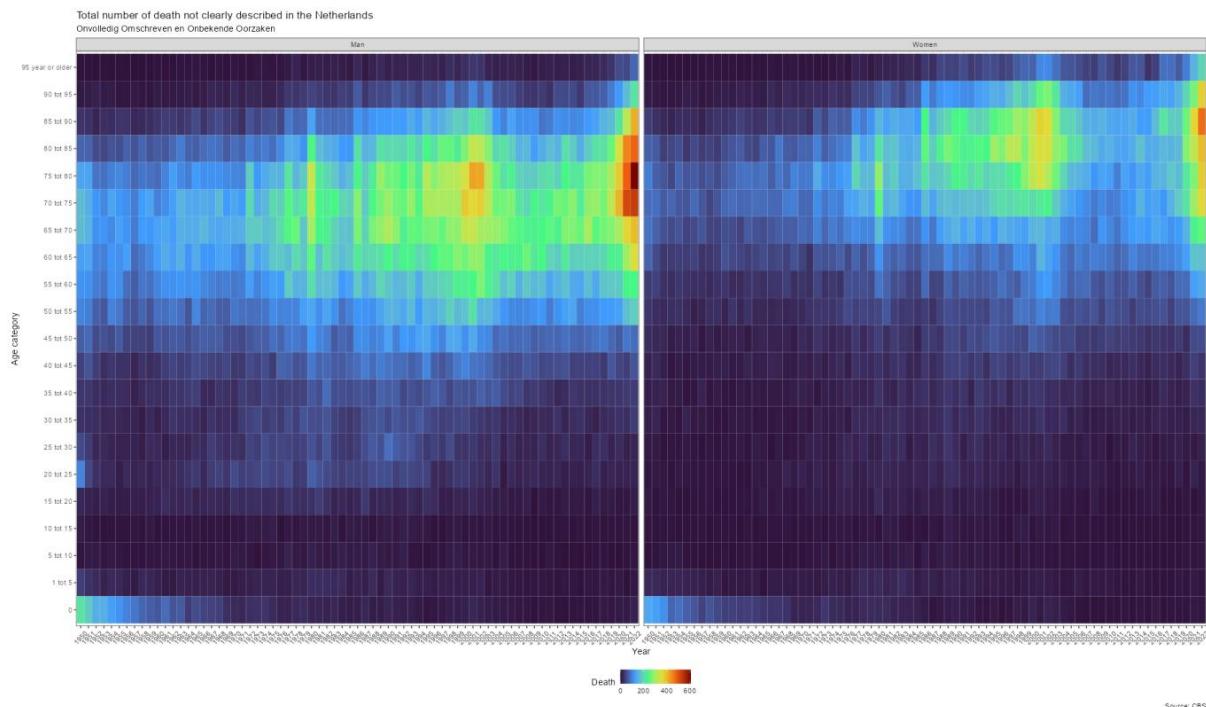


Figure 160. Total number of deaths not clearly described per year, sex, and age category.

Hence, in the end, we do believe that there was excess mortality (Figure 157), but that the exact number is just a function of the underlying model. It is better to look for possible scenarios that may have led to the observed (and thus excess) mortality.

The excess mortality from 2020 and 2021 can be fully explained by Covid-19.

Although this hypothesis is not the main aim of our research it is a hypothesis that we cannot bypass. In our search for open data, we found several data sources provided by CBS that look for specific [causes of death](#) (fig. 161). Some of these sources particularly addressed the rate of death attributed via [Covid-19](#) in which a distinction was made between confirmed Covid-19 death and plausible death.

Before we dive deeper into this scenario based, we must first address the issue of reporting dead people. In the Netherlands, a physician needs to fill out a B-form (cause-of-death form).¹²⁸ The B-form has two parts: the first part describes the chain of events that have led to the recorded death (from direct cause of death to underlying cause of death). In the second part the physician has space to describe factors that may have come contributed to the recorded death but are not part of the immediate chain of events. For instance, the illness that directly led to the death of a Dutch citizen could be acute respiratory distress syndrome (ARDS) which was caused or the cause of pneumonia which was caused or the cause of a suspension of Covid-19. This person also had high blood pressure and was a diabetic.¹²⁹

¹²⁸ <https://www.cbs.nl/nl-nl/deelnemers-enquetes/decentrale-overheden/overzicht/doodsoorzaakverklaring>

¹²⁹ <https://www.cbs.nl/nl-nl/deelnemers-enquetes/decentrale-overheden/decentrale-overheid/doodsoorzaak/adviezen-cbs-gebruik-van-covid-19-op-doodsoorzaakverklaring--b-formulier-->

All B-forms are collected by CBS, but one can easily imagine that this takes quite some time if done manually. Also, it is very important that standardized ways of reporting and coding are applied, for instance by using ICD codes. Therefore, in 2013, CBS changed to an automatic coding system called IRIS.¹³⁰ IRIS is an interactive tool for which multiple causes of death can be coded, and the underlying cause of death is selected according to the rules and guidelines published by the WHO in the ICD classification.¹³¹ However, the change in coding system also led to a change in the top five causes of death.

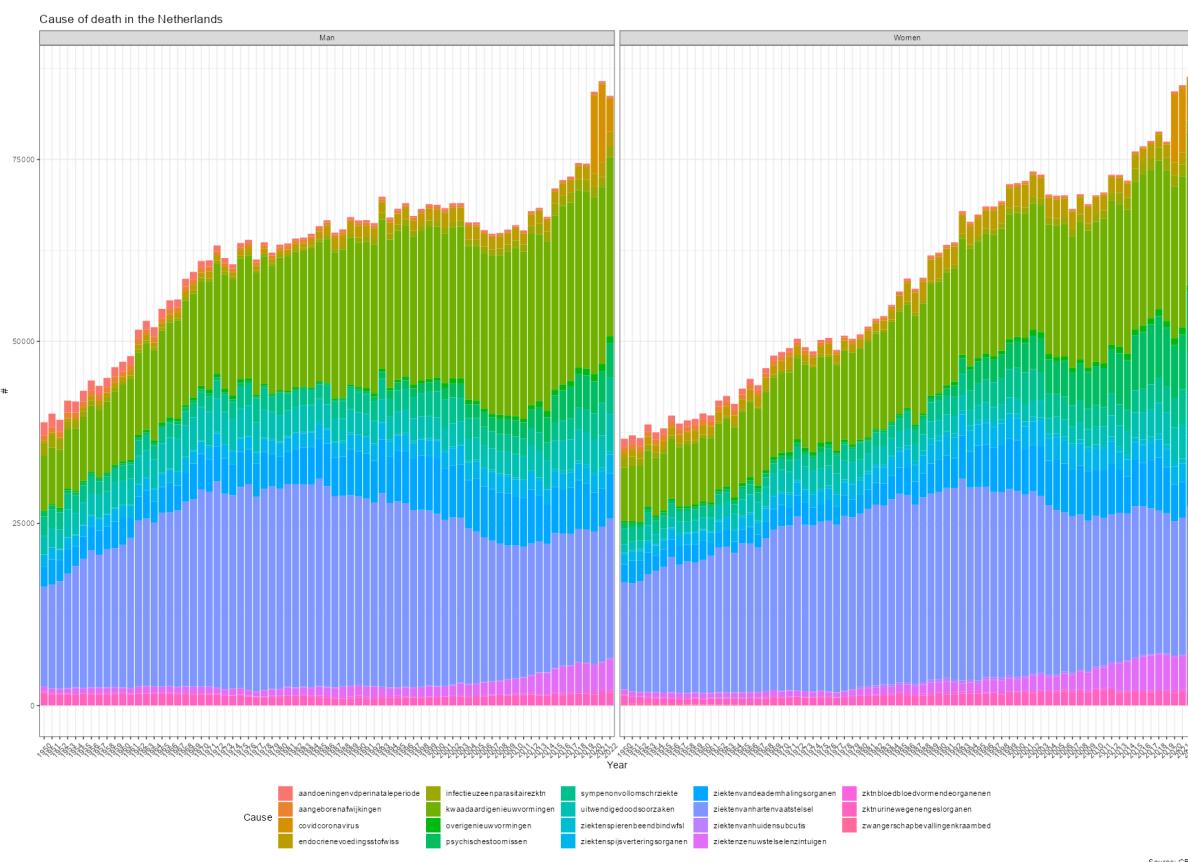


Figure 161. Cause of death (global categories) per year, and sex.

During the Covid-19 pandemic the manual variant was re-introduced because Covid-19 was not part of the current ICD-10 coding system and the WHO guideline changed with respect to Covid-19. On the April 20th of 2020, the WHO published the following with regards to the registration of Covid-19 attributable deaths:¹³² “A death due to COVID-19 is defined for surveillance purpose as a death resulting from a clinically compatible illness, in a probable or confirmed COVID-19 case, unless there is a clear alternative cause of death that cannot be

¹³⁰ <https://score.tools.who.int/tools/count-births-deaths-and-causes-of-death/tool/iris-automated-coding-system-for-causes-of-death-97/>

¹³¹ https://www.cso.ie/en/media/csoie/methods/birthsdeathsandmarriages/Information_Note_-_IRIS.pdf

¹³² <https://www.cbs.nl/en-gb/news/2022/42/1-420-covid-19-deaths-in-q2-2022/cause-of-death>

related to COVID disease (e.g. trauma). There should be no period of complete recovery from COVID-19 between illness and death. A death due to COVID-19 may not be attributed to another disease (e.g. cancer) and should be counted independently of preexisting conditions that are suspected of triggering a severe cause of COVID-19.” In short this means that any death in which Covid-19 was suspected to be present would most likely be deemed the underlying cause of death. This is a clear deviation from the way previous infectious diseases had been handled.

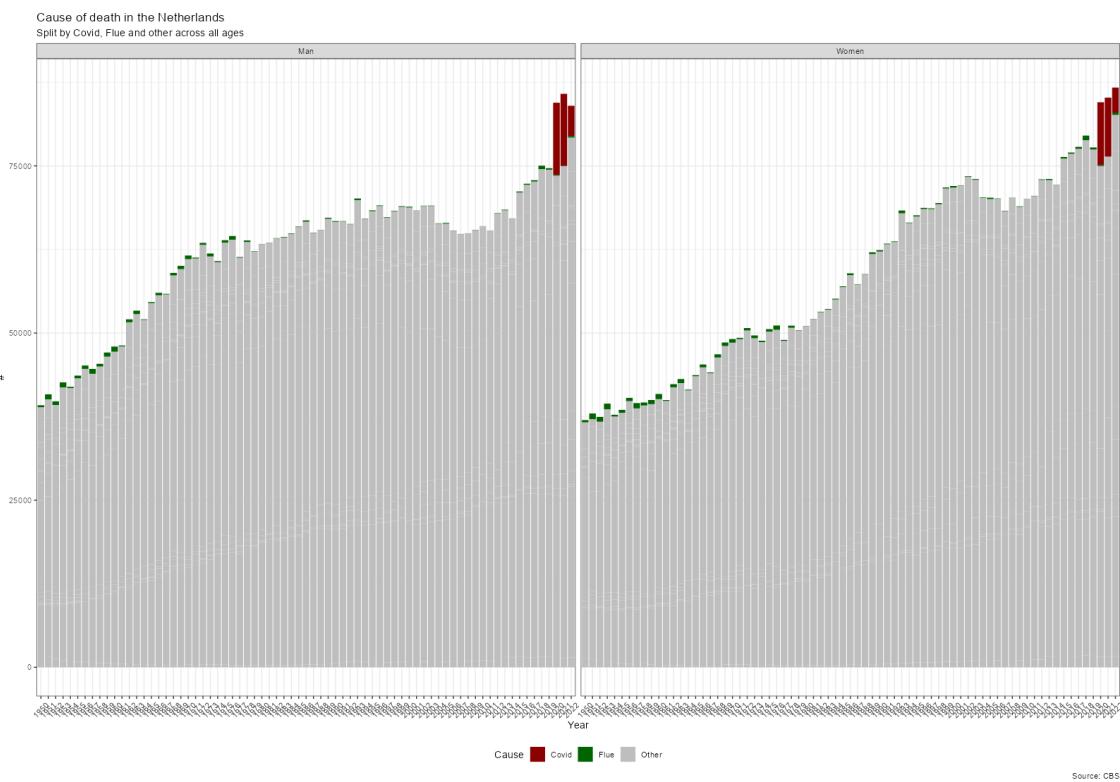


Figure 162. Cause of death split by Covid-19, flu and the rest per year and sex.

This last remark can be clearly shown by comparing the number of deaths registered to be the cause of Covid-19 or the flu and by comparing the all-cause mortality numbers across the years (Figure 162).

We remark that the numbers published by CBS are questionable. To be more precise: The RIVM and the ECDC report a total of around 23,000 deaths due to Covid-19,¹³³ while the CBS reports 48,000. Thus, the numbers cannot immediately be trusted at all.

Although 2018 shows a clear peak in mortality compared to the previous three years, there seems to be no strong increase in flu-attributed deaths. In fact, when looking at the CBS data for causes of death, the flu has not been a big contributor on death for the past 40 years. However, Covid-19 is a big contributor to the mortality when following the statistics

¹³³ <https://www.ecdc.europa.eu/en/publications-data/data-national-14-day-notification-rate-covid-19>

surrounding the causes of death. A possible way to look closer, despite being somewhat circumstantial, is to connect the metrics on all-cause (total) mortality, excess and expected mortality and Covid-19 mortality (Figure 163). For connection purposes, data needed to be aggregated to a higher (monthly) level.

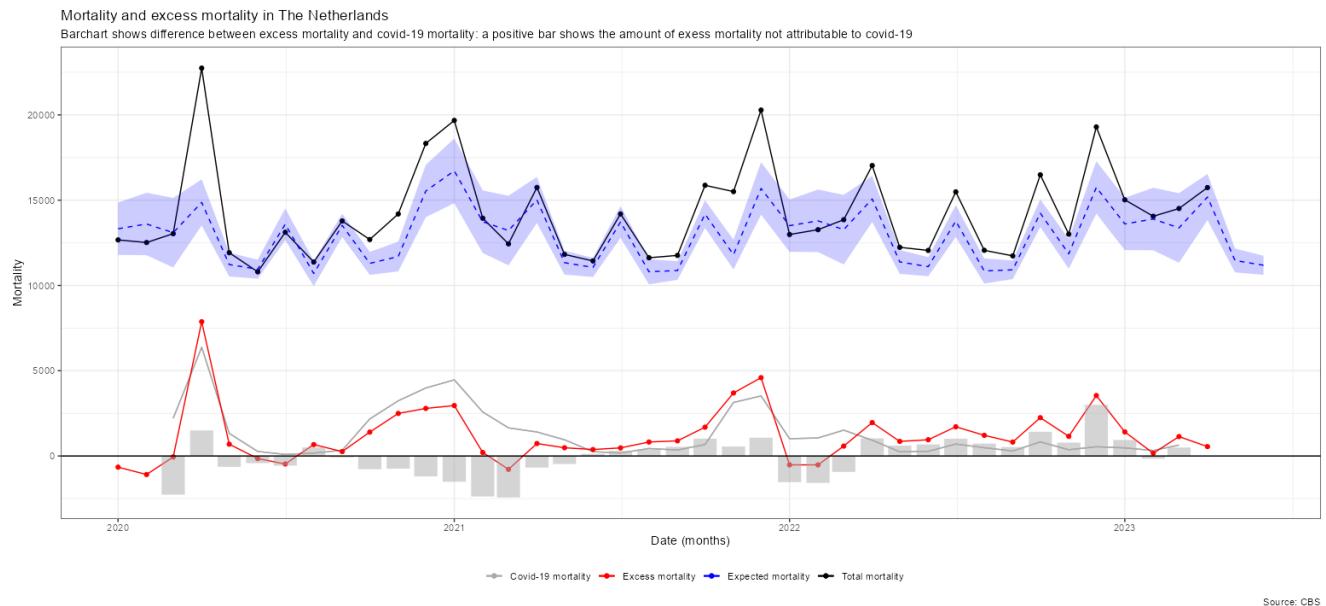


Figure 163. Observed, expected, excess and Covid-19 mortality. The bar chart shows the difference between excess mortality and Covid-19 mortality.

In the figure above, which only shows data from 2020, the peak in mortality in March 2020 can be clearly seen. Regardless of cause, this is observed total mortality and so must be treated as ‘the truth’ because somebody is either dead or alive. Furthermore, we can see how the expected mortality follows somewhat of a traditional periodic curve which now looks a bit strange since we had to summarize the data from weeks to months. Nevertheless, one can clearly see when the observed mortality exceeds the expected mortality, which leads to excess mortality. In addition, we added Covid-19 mortality and the difference between Covid-19 mortality and excess mortality. That difference shows how much of the excess mortality can be contributed to Covid-19 mortality considering all other things equal.

From 2020 to 2023, there are many months that show excess mortality. This is thus unexpected mortality. For almost all of 2020 and the beginning of 2021, the Covid-19 mortality exceeds the excess mortality until Covid-19 mortality decreases. From that point on the excess mortality number is higher than the Covid-19 mortality number. So, what is happening here? Well, Covid-19 is a new disease so it is hard to believe it would be expected. Therefore, deaths attributed to Covid-19 are almost immediately ‘excess’ deaths since Covid-19 was never factored in. So, if we see the excess mortality curve in a form that is parallel to the Covid-19 mortality curve, it makes perfect sense even if the amount is not

the same (there could always be other fluctuating unexpected causes). Here, the sudden increase is most likely Covid-19.

But what about the period in which Covid-19 mortality does not exceed excess mortality? Or the curves are no longer parallel? Well, in a month in which Covid-19 mortality does not exceed excess mortality, the excess cannot be fully attributed to Covid-19. Once again, the inherent limitation of the excess mortality metric is its calculation which is based on expected mortality which is a demographic model based on year of birth, age, sex, and smoking status. Seasonal patterns are not even modelled by specific infectious diseases but by the cycles of the seasons themselves. That is why, for a dynamic model, the model is still quite stable as can be seen in the graphics showing the expected mortality (Figure 156). Thus, if the Covid-19 mortality is lower than the excess mortality, in a time in which countermeasures such as social distancing are still in place, it must mean that other causes of death have taken the lead (or the countermeasures were not effective).

What does all this mean for our prior comments on the reporting of plausible (Covid-19) attributable death? Well, it could mean that Covid-19 was favored as a cause of death. Although it is clear to see that the all-cause mortality in 2020 and 2021 was much higher than the years before, it is perhaps too easy to contribute them all to Covid-19 when Covid-19 was suspected. This is especially clear for 2021 (and even 2022) in which all-cause mortality was much higher than the years before, yet the numbers of Covid-19 attributed deaths decreased.

In this discussion, we did not take another effect into account, namely that excess mortality in one time interval would naturally be followed by lower than expected mortality later. This is something that has been visible during and after flu-epidemics. This makes the whole discussion about excess mortality even more complicated.

Does this mean that this opens the doors for a case in which death can be attributed via delayed cancer care? Before starting to answer the next two hypotheses one must first look at the number of deaths attributed to cancer. From the graph below, it is clear to see that cancer deaths have been rising for years although the past five years showed a small decline in males and a steady level in females (Figure 164). No significant drop can be seen for 2020 and 2021, and a small increase can be seen for 2022. However, we cannot rule out (based on ecological data), that the reporting on the causes of death is ‘contaminated’ by the times in which they were done meaning that from 2020 – 2022, since almost every aspect of life was affected by Covid-19.

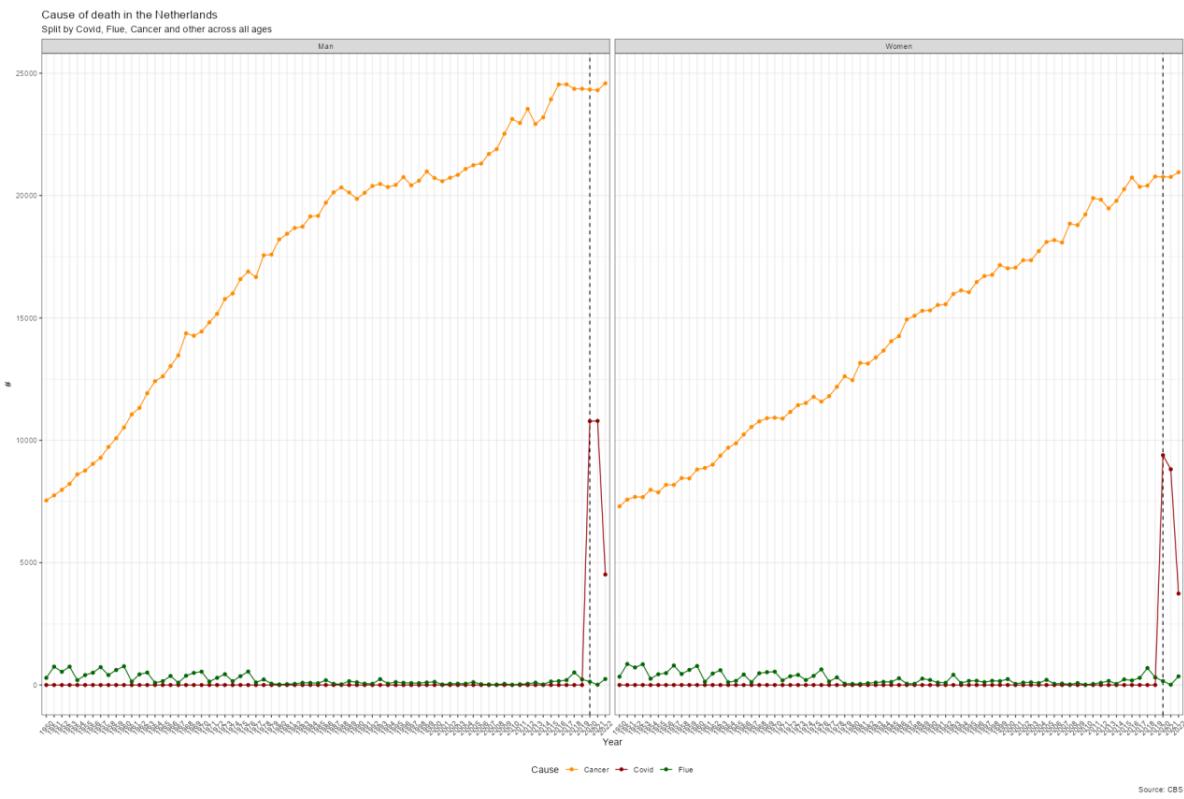


Figure 164. Cause of death split by Covid-19, flu and cancer across sex and inclusion year.

Last, but not least, it is important to mention that the numbers of death that did not have a particular cause attached to it, did increase in the past three years, and saw numbers much larger than have been seen in the decades before.

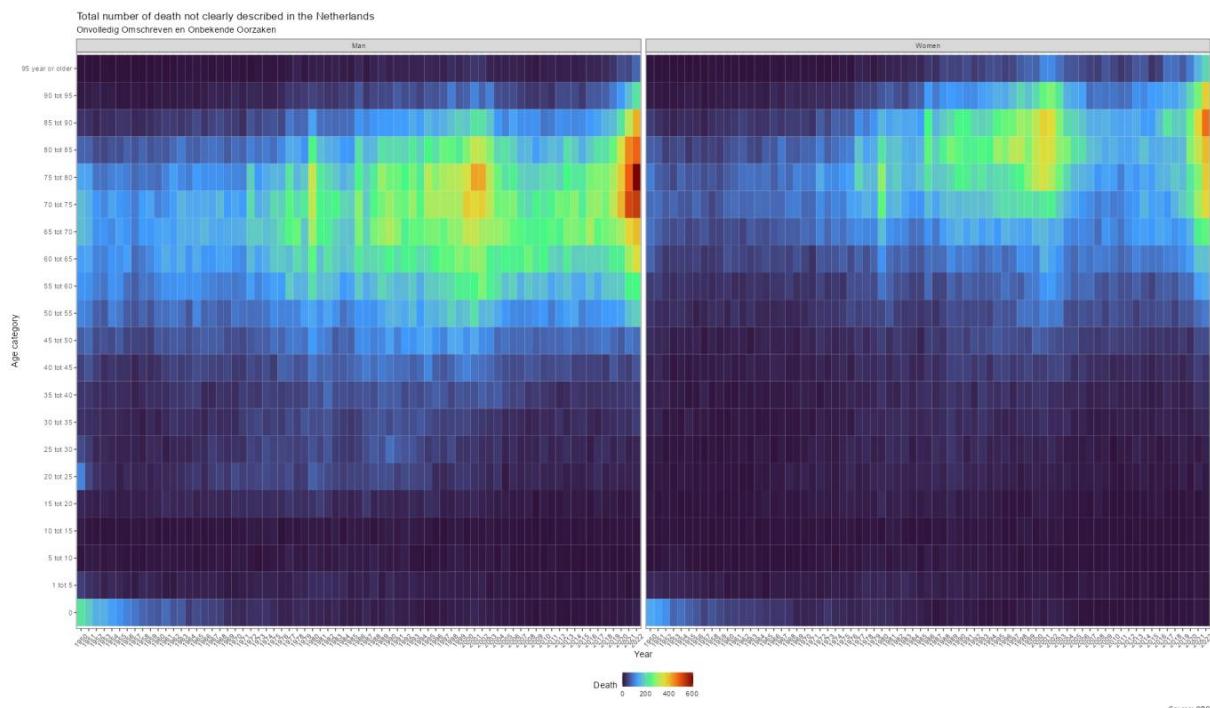


Figure 165. Total number of deaths not clearly described per year, sex, and age category.

In summary, we believe that it is impossible to disentangle how many of the deaths counted in 2020 and 2021 were directly caused by Covid-19, especially when the biological pathway and the disease itself was still debated. There is also a clear difference between the number of deaths attributed to the flu and to Covid-19 despite previous seasonal peaks in excess mortality.

Nevertheless, certain peaks are clearly caused by Covid-19 and Covid-19 did play a major role in the mortality observed for 2020. However, for 2021, the link seems far less clear by a decreasing Covid-19 mortality. It is here, perhaps, that cancer could play a role via the pathway of delayed care (Figure 166).

The excess mortality from 2020 and 2021 can be (partly) explained by the temporary cessation of population screening.

The possible influence of delayed screening on mortality, and thus on excess mortality (since it is the first-time screening has been halted), is all about time. The sole purpose of screening is to detect cancer before symptoms arise in which case the tumor has most likely progressed to a stage of lower survival. By being able to detect cancer earlier, it should be possible to give the patient a higher degree of survival.

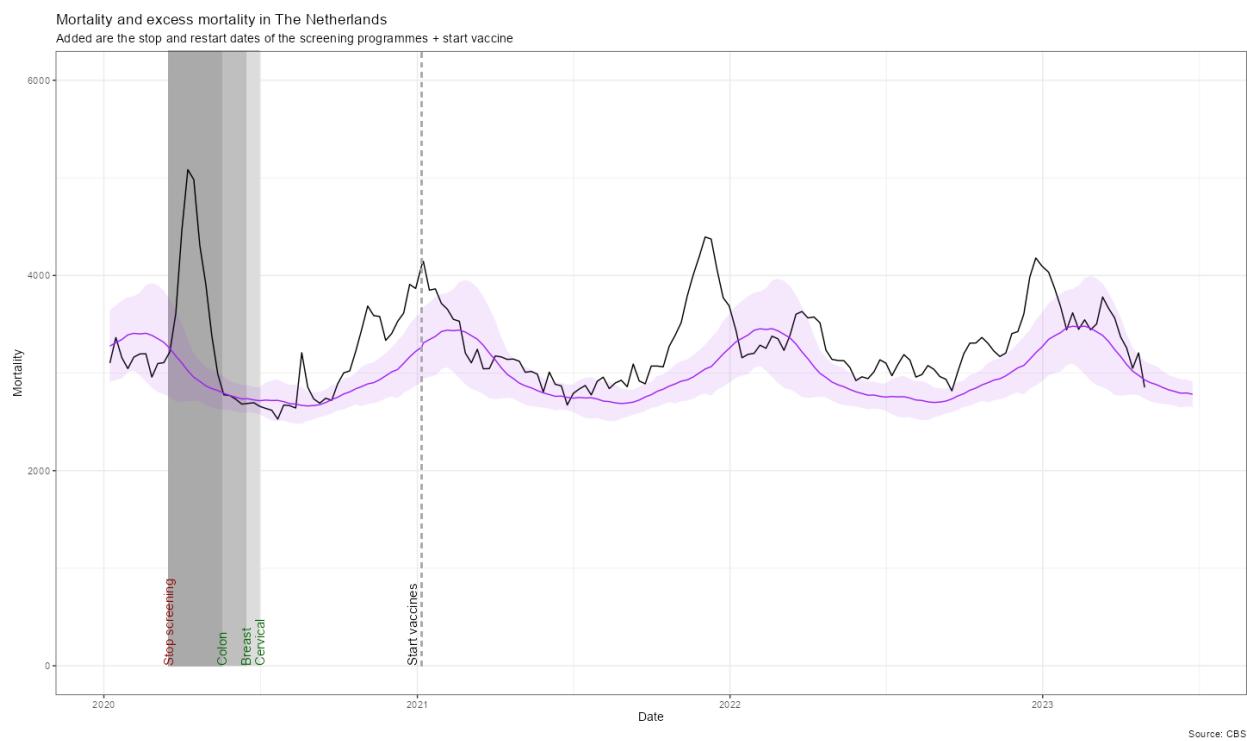


Figure 166. Observed mortality (black line), expected mortality (pink area), and excess mortality. Also shown are the dates at which screening stopped, was resumed for colon, breast, and cervical cancer, and when the first round of vaccines started.

The exact gain of screening is, however, not without discussion. For instance, people who are diagnosed earlier may seem to survive longer but the gain in survival is possibly not a real gain, but an increase in time being diagnosed. This phenomenon is called *lead-time bias*.

The matter is not quite settled scientifically, and it would be hard to argue that no one will benefit from screening. Cancer is a disease in which cells spread uncontrollably and do damage to the human body. These cells integrate into tumors and can spread via the blood stream or the lymph nodes. The earlier a disease is detected, and thus treated, the higher the probability that the disease is stopped from spreading. Once metastasized, cancer is much more difficult to treat to the extent that curative options are no longer available.

From the open data that we analyzed; our findings are in line with what [has been reported all over the world in which Covid-19 caused a pause in healthcare](#). That is: screening rates declined which is a straightforward result of not doing any screening. In addition, the number of diagnoses made also dropped, which is another straightforward consequence of not doing any screening. For all three types of cancer (breast, cervical, and colon) large scale screening initiatives exist. When they stop, a steady stream of cancer diagnoses also stops. The main question we needed to ask was not if the screening and diagnostic rates declined (it is easy to deduce that they would) but to determine if the excess mortality rates of 2020 and 2021 could be partly explained by the decrease in diagnostic rate. In other words: would a three-month delay in screening lead to such a drop in cancer survival that such a person would not have died in 2020 or 2021 had he or she been diagnosed via screening during the screening stop?

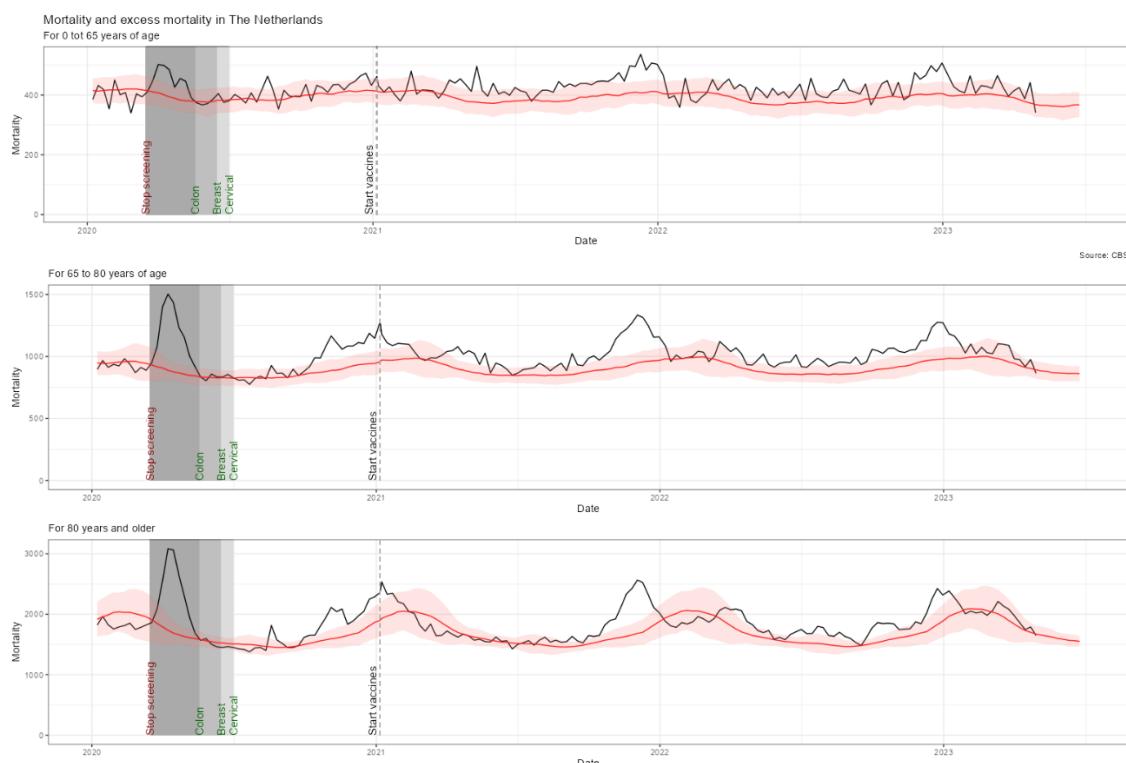


Figure 167. Observed mortality (black line), expected mortality (red area), and excess mortality. Also shown are the dates at which screening stopped, was resumed for colon, breast, and cervical cancer, and when the first round of vaccines started.

First, and foremost, we are not able to answer such a question on an individual level: we do not have individual level data. Hence, we cannot see when a patient was scheduled for screening and when a patient eventually underwent screening. We also cannot see if such a patient is still alive. A most straightforward analyses would have been to collect all patients who died in 2020 and 2021 and see if they had been diagnosed in 2020 or 2021. We would then have been able to see what their tumor stage and tumor grading would have been and from that we would have been able to deduce their survival probability. However, we do not have such data, so we can only try and connect ecological data from different available sources to the findings reported in literature to the clinical expert reviews.

The literature on the impact of halted screening on cancer mortality is almost unanimous in their writing that the delay in cancer screening may have an impact at some point. However, none of the articles we have read mention an immediate impact for breast, colon, or cervical cancer. The reason for this is that the types of cancer that are detected via screening are mostly those without symptoms at an early stage. Most of these stages have a very good five-year survival rate. This is especially true for breast cancer, which is by far the largest group of cancer patients. Hence, not detecting as many Stages I or II tumors may not necessarily mean that their survival is impacted. Had the delay in screening been much larger than three months (say one year) the impact would most likely have been much larger.

The data we collected clearly showed a decrease in screening rates, but not an increase in death. Death rates for each of the cancer groups, across ages and stages, did not show a reversal in trend. Once again, we are not able to look at individual cases, and so it might very well be that somebody might have lived beyond 2021 had they been detected earlier, but it is very unlikely that such a person would have a completely different survival curve. This is because all patients that could not be screened at the beginning of the pandemic still received an invitation to get screening done once screening restarted. Therefore, for each possible new cancer patient, the delay in care is around three to four months. Considering the stages usually detected via screening, it is unlikely that this would have an impact on survival. When we calculate the survival curve from a theoretical breast cancer patient that has either Stage I or II with the number of positive lymph nodes mostly seen in these patients, it clearly shows that their five-year survival is very good. It is therefore unlikely they would die in the first two years. The PREDICT model, which includes screening as a factor, also shows that screening only adds 1% to the survival rate.

Another possibility, which we could not check explicitly, was fear. It is possible that a person eligible for screening, still postponed screening out of fear for contracting Covid-19 in a hospital. However, should a person be detected a year later because of fear of screening, in a time where screening is available, the drop in survival is not due to delayed care but because of the choice made not to undergo screening.

The excess mortality from 2020 and 2021 can be (partly) explained by delayed care of already diagnosed cancer patients.

From what we have [read in the literature](#), and analyzed on the [OpenDIS and VEKTIS data](#), it seems highly unlikely that delayed care of already diagnosed patients explains the excess mortality. And indeed, across none of the figures we produced did we find a trend reversal like the one we saw for screening and diagnosis.

For total expenditure we did find a small trend reversal of increasing costs (Figure 70), but when we looked at the DBC coding it did not seem that this affected any of the treatments normally conducted for these three groups (Figure 100). In essence, we did not see large scale changes when we ranked the treatments for breast (Figure 96 and 98), colon (Figures 103 and 104), and cervical cancer (Figures 107 and 108). We also did not see notable changes when we split the number of DBCs per healthcare provided (Figure 97, 100 and 105).

This of course does not mean that things do not change, and especially the graphs ranking the treatments show changes across the years. This makes it especially difficult to look for small trend reversals but relatively easy to find large ones – in essence, you would expect to see large scale changes like screening and diagnosis. What we did see, however, were things like a shift in location for chemotherapy (from hospital to home – Figure 96).

In the end, and like all the other things we have seen in both the literature and our data with regards to delayed care for breast, colon and cervical cancer, we could not find a trend reversal that would indicate that diagnosed patients did not receive their much-needed care.

Of course, had we received the data from IKNL from which we would have been able to derive patient-level treatment trajectories (including time-from and time-to), we would have been able to model the importance of delayed care. In fact, we would have been able to provide some insight on what ‘delay’ would mean clinically. Nevertheless, as stated many times before, we did not receive the data, but we do believe that a large part of the answer lies within the granularity of those datasets.

Limitations of our research

The limitations of our research are reflected by the granularity of the data that was made available to us. When ecological data is all that one has, the only real methodological sound options that you have is look for trend reversals (sudden increase or decreases in causes of death) and examples of co-integration (a change in factor A leads to a change in factor B but not vice versa). For most of our work, this is what we have been trying to do paired with the findings of published research and prediction models.

The most important factor of our research question is time - meaning time at the patient level. If delayed care truly has an influence on patient survival, either by being diagnosed too late or by being treated too late (or both) this should be reflected by an increase in time from invitation to diagnosis, or from diagnosis to treatment. If a patient would have a higher than expected tumor stage, which would mean a tumor stage via screening which deviates from the norm, there would be some reason to assume that perhaps an earlier screening could have made a difference. However, there are currently no statistical models that model the clinical pathway of disease progress. Hence, it is almost impossible to say for a given patient how fast her tumor is spreading and there is clear variation within tumor stages and tumor grades.

Our analyses are clearly not robust enough to pinpoint cause, nor to provide a definitive answer to the question how many patients considered excess deaths are dead because of delayed care. We also cannot say anything about the difference between 2020 and 2021. The data simply do not allow estimates at the patient level from which a more ecological number can be presented and then compared with the recorded excess mortality.

Furthermore, our two rapid reviews are indeed rapid. We did not apply a full systematic approach including risk of bias assessments because we wanted to only gloss over the findings of others to look for the main narrative. That main narrative is that breast cancer mortality is influenced by specific variables such as tumor stage and grade, and the number of positive lymph nodes. In addition, it seems to matter if certain markers are present or not which is reflected in the breast cancer prediction models we used. Furthermore, it seems that most of the literature agrees that the delay in care must have some impact on cancer mortality, but that the impact is not immediate. The specific size of the impact also remains to be determined but most authors agree that three months of delayed cancer care via screening and actual care has limited influence on cancer mortality. Of course, it may very well be that we missed important papers, but none of the 70+ number of papers on the impact of delayed care on mortality mentioned a direct and immediate increase in mortality.

The use of validated prediction models combined with the patient population from the IKNL synthetic breast cancer data set is of course also not without limitations. By using validated prediction models, which are routinely used in cancer care, we tried to discern what the predicted survival rate would be for the population most often seen. Once again, this is an ecological approach, and not a patient-by-patient approach, which would have contained much more information. The limitations of the synthetic data itself has also been made clear multiple times, even to the extent where we perhaps could have just decided to forfeit the exercise.

Directions for future research

Future research on this topic must include patient-level data. We believe that for The Netherlands such data is already available at the IKNL. We also believe that the statistical analysis plan we drafted for that data will be sufficient to determine if delayed screening and/or delayed care will play a role in cancer survival. New prediction models on this topic must include time from invitation to actual diagnosis by screening, and from diagnosis to care. None of the current models include these features, most likely because care trajectories are so standardized. In that regard, Covid-19 is a great catalyst for including these time-metrics especially in a world in which the pressure on health care is mounting due to an ageing population and a decline in health care workers.

Final thoughts

This report would have looked completely different had we been given access to the patient-level data we deemed necessary to provide more robust inferences. Nevertheless, in the absence of such data and the presence of only openly available data, it seems highly unlikely that the excess mortality seen in 2020 and 2021 is due to delayed health care. Not only are there no hints for such a mechanism in the literature, but there is also no indication for it in the data openly available, as we have shown in detail. We did not observe any serious trend reversal apart from the decrease in screening rates and cancer diagnoses. This means that the excess mortality seen in 2021 and beyond, which is higher than the Covid-19 mortality, must have a cause (or multiple causes) somewhere else.