UiT Norges arktiske universitet

Using real world data for pharmacoeconomic assessments - a case study

Jeroen Martin van Zuiden Haukaas

May, 2024

Table of Contents

Preliminary Content

Acknowledgements

I want to thank a few people that have contributed to this thesis.

I extend my gratitude to my advisors for their invaluable contributions to this thesis. My sincere thanks to Lars Småbrekke for his crucial guidance and insightful discussions on statistical methods. I appreciate Gro Live Fagereng for her clinical expertise in reimbursement systems. Lastly, I’m grateful to Helle Nærsnes Endersen for her counsel in pharmacoeconomic methods and innovative approaches. I also want to thank Yngvar Nilssen for providing insights regarding the cancer registry. I am also grateful to my advisors for enduring the initial drafts of this thesis, a task arguably as challenging as the research itself.

I want to thank Erik Sagdahl for not only suggesting the topic of this thesis but also for enabling me to stay at Sykehusinnkjøp HF during its development.

I am also thankful for Sykehusinnkjøp HF generously covering the expenses for the procurement of the registry data.

I want to thank my family and Modesta for support and motivation while working on the thesis.

Preface

This master thesis was composed using the R bookdown package, specifically the “thesisdown” package by GitHub user ismayc, which I have adjusted to align with the University of Tromsø’s format requirements.

The work presented here aims to explore the real-world cost-effectiveness of pembrolizumab monotherapy for stage IV NSCLC patients. It represents the result of research conducted between September 2023 and May 2024 at the University of Tromsø, as part of the clinical pharmacy and pharmacoepidemiology research group (IPSUM).

Abstract

Pembrolizumab monotherapy for non-small cell lung cancer (NSCLC) was deemed not cost-effective due to limited data, high uncertainty regarding long term results and a considerable budget impact. However, due to confidential price agreements it was approved in 2017, as first-line treatment for metastatic NSCLC that expresses PD-L1 ≥ 50 %.

Given potential discrepancies between clinical efficacy and real-world effectiveness due to differing treatment guidelines and inclusion criteria, it's uncertain whether the treatment is truly cost-effective, especially since projections are based on immature data.

Our objectives are to (1) re-evaluate the economic feasibility of pembrolizumab in treating lung cancer within the Norwegian healthcare framework with real world data; (2) to identify the possible variables contributing to the gap between its efficacy in controlled studies and its effectiveness in a clinical setting and (3) to provide a framework for assessing the cost-effectiveness of treatments with immature data.

We included real-world data from 2016-2022 in the Cancer Registry of Norway (CRN).The clinical data includes 3592 patients, with 3642 cases of lung cancer, and we estimated mean and median progression free and overall survival using Cox regression and Kaplan-Meier. These results will be compared to a digitized version of the KEYNOTE-024 study. A new pharmacoeconomic assessment was conducted, similar to the original assessment.

There was a significant difference between real world data and clinical studies (p = 0.001). Median survival for the real world participants was 13 months (95% CI 12-15) and 26.3 months (95% CI 19.4-41.4) for patients in the KEYNOTE-024 study. However, when adjusted for ECOG status there was no significant difference in OS (p = 0.07). The treatment had an ICER range of 474 621 - 689 641 NOK depending on disease progression.

Although there are several limitations to real-world data, it can be a useful tool in assessing real world effectiveness when treatment guidelines will differ from clinical trials. Pembrolizumab had a significant efficacy-effectiveness gap mostly due to differences in ECOG status, given the limitation of the results the treatment could still be considered cost-effective.

1 Background

1.1 Healthcare priorities

The Norwegian parliament has defined three main criteria for prioritizing treatments in healthcare. The criteria are utility, resource cost, and severity of the disease. These criteria are assessed as a whole and balanced against one another. Judgement calls are to be included in a comprehensive consideration of the measures. This especially pertains to considerations of uncertainty in the evidence and the cumulative impact on the budget. (1).

The criteria for utility dictate that the priority of an intervention increases with its anticipated health benefit. The anticipated utility of an intervention is assessed based on clinical evidence of increased life expectancy, quality of life, reduced function loss, physical or mental improvement, and reduction of mental or physical pain or discomfort (1).

The resource criterion stipulates that the priority of an intervention increases as it consumes less resources (1).

The severity criteria dictate that the priority of an intervention increases with increased disease severity. The severity of a condition is based on risk of death or function loss, the degree of function loss, pain, and mental and/or physical discomfort. The severity is based on the current condition, and on the prognosis (1).

The treatment of NSCLC is classified as a condition of high severity, and treatment with pembrolizumab has a significant financial impact on the budget. In the initial health technology assessment (HTA) MSD the estimated cost is 133.3 million NOK yearly with maximum price but according to the Norwegian medical products agency (NOMA) the cost is closer to 550 million NOK, but the estimates are uncertain. The large budget impact combined with uncertain estimates for utility further complicates the pharmacoeconomic assessment (2).

1.2 Nye metoder (New methods)

The Norwegian national system for prioritizing medical methods or interventions to be funded by the regional healthcare providers (RHF) is called Nye Metoder. It was established in 2013 with several key objectives:

1. To assure patients that the new method has been assessed for effectiveness and safety.

2. To support equal and swift access to new and innovative treatments.

3. To demonstrate the added benefits of new treatments compared to existing treatments.

4. To ensure that treatments no longer considered satisfactory are phased out.

5. To generate a quality-controlled framework for decision-making, prioritizations, and resource management.

6. To offer transparent decisions (3).

1.2.1 The reimbursment system for pharmaceuticals

The Norwegian reimbursement system has four main phases. The first phase is the proposal phase, where anyone can submit a proposal for a new treatment method. The proposal will be considered by the Ordering Forum if it is a pharmaceutical treatment, and they will prioritize which pharmaceuticals will undergo a health technology assessment (HTA) (4,5).

The second phase is assessment of the method ordered by the Ordering Forum. There are multiple types of assessment methods within the new methods system based on the intervention and use-case. In this thesis the assessment is a rapid HTA for a drug. It focuses on efficacy, safety, and cost-effectiveness. A rapid HTA is mainly based on documentation sent in from the drug producer or supplier. After receiving the documentation NOMA has 180 days to perform necessary analyses and deliver their assessment (6).

The third step involves reviewing the provided assessment, and the Decision Forum then decides whether the method can be implemented in the specialist health service (hospitals). The final step is implementation, which involves adjusting current treatment guidelines (7).

1.2.2 Current challenges with HTAs in Norway

Although NOMA has a 180-day deadline for evaluating new treatments, the average processing time extends to 374 days in 2022. However, when accounting for the waiting period for receiving documentation from pharmaceutical companies and the capacity of case handlers, the effective processing time averages 228 days (8).

This is cause for concern, since delayed implementation of treatments will negatively affect patients in Norway, who may have to resort to inferior treatments due to the prolonged processing time. The reimbursement system aims to reduce processing time and has created a fast-track system for PD1/ PD-L1, which will expediate the processing time significantly for these treatments. In turn it will also reduce the waiting time for other pharmaceuticals (9).

HTAs implemented temporarily or with alternative price agreements tend to have a long processing time; usually around 250 days (10,11). It is worth noting that processing time might reduce when issues with the system are resolved and case handlers are more experienced with the system.

1.2.3 Temporary implementations

There are multiple treatments with temporary implementation in Norway, but the exact number is unclear due to varying terminology. For example, Rozyltrek was temporarily approved with an alternative price agreement (10). In other cases the temporary implementation is called conditional implementation, which often include a required re-evaluation with updated data after a certain time (11,12).

The distinction between these types of approvals remains unclear, potentially due to inconsistent use of terminology. This inconsistent use of terms is possibly due to the fact that only three treatments have been through the conditional/temporary approval (13).

1.2.4 Alternative price agreements

A framework for alternative price agreements allows a unique agreement for a specific treatment without creating precedence for other cases. These agreements should be as simple as possible, and not contribute to a considerable increase in administrative workload. The stipulations in the agreement will also be public, and the uncertainty regarding price and budgetary impact shall be limited. Follow-up, data sources and other practical issues must be clarified, and all terms of the agreement must also be in accordance with the prioritization criteria (14).

Alternative price agreements aim to enhance patient access to innovative and effective treatments where the usual flat price reductions are unsuitable. This scenario frequently applies to newly introduced, costly treatments with immature data. There are less than 20 treatments approved with this method since the framework was developed in 2020 (13,14).

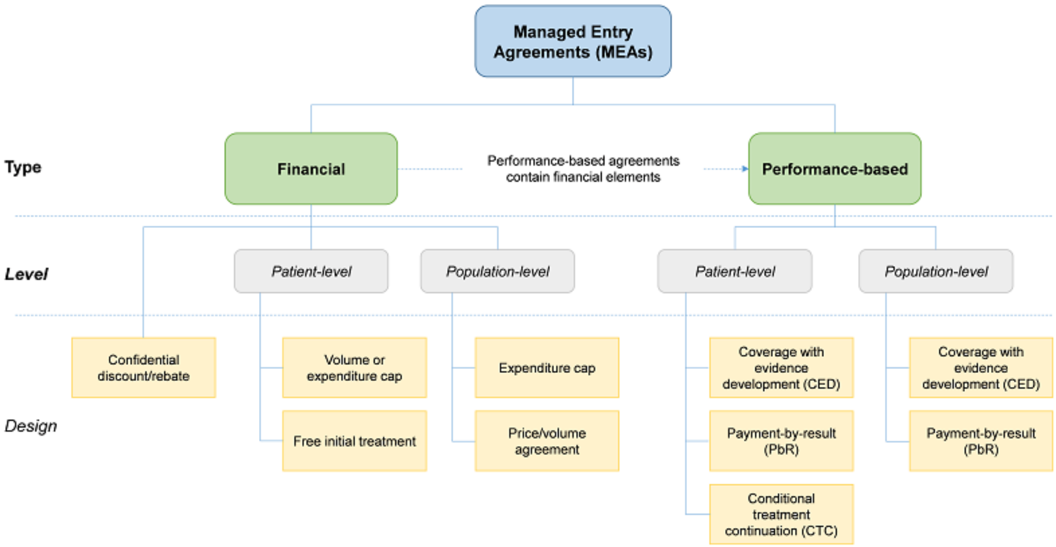

Managed Entry Agreements (MEAs)
Type
Financial
Performance-based agreements contain financial elements
Performance-based
Level
Patient-level
Population-level
Patient-level
Population-level
Design
Confidential discount/rebate
Volume or expenditure cap
Free initial treatment
Expenditure cap
Price/volume agreement
Coverage with evidence development (CED)
Payment-by-result (PbR)
Conditional treatment continuation (CTC)
Coverage with evidence development (CED)
Payment-by-result (PbR)

Figure 1.2.1: A overview of the types of alternative price agreements which are possible. The simplest agreement would be a confidential discount (15).

7 out of the 14 treatments implemented with alternative agreements seem to be linked to a confidential discount, three are linked to volume, and two to performance (13).

1.3 Advances in lung cancer treatment in Norway

Recent advancements in immunotherapies have reduced morbidity and mortality for lung cancer patients. The 5-year survival rate has nearly doubled during the last 20 years, reaching 26.5% for men and 33.7% for women. In 2022 there was 3524 incident cases of lung cancer in Norway, which is estimated to increase to 4000 cases in 2030 (16).

1.4 Findings in HTA of pembrolizumab for metastatic NSCLC

Pembrolizumab monotherapy was approved as first line treatment in Norway in 2017 for metastatic NSCLC expressing PD-L1 ≥50 %. The Norwegian medical products agency (NOMA) estimated that on average the new therapy would increase quality adjusted life years (QALYs) by 0.98 and increase life expectancy by 1.19 years. The basis for these estimates is data from the KEYNOTE-024 study. These estimates were more conservative compared to the manufacturer's projections, which were 1.01 and 1.26, respectively. These estimates were based on the KEYNOTE-024 trial, where median follow-up was 11.2 months. NOMA noted that there are high levels of uncertainty around long term effects due to the short follow-up and this complicates the pharmacoeconomic assessment (2,17).

Metastatic NSLC was recognized as a serious disease, with an estimated loss of 11.5 QALYs. Therefore, the disease has high severity, which increases the priority of the treatment. The utility and resource criteria were challenging to estimate due to the dataset, which was of high quality but short follow-up (1,2).

The budget impacts of implementing the treatment were substantial and estimated to cost 500 million NOK, which de-prioritizes the treatment regarding the resource criterion. However, NOMA did consider the treatment cost-effective, but not for the public price (1,2).

Recently an update of the KEYNOTE-024 was published with a follow-up of 5 years. Median overall survival (OS) was 26 months in the pembrolizumab group, compared to 13.4 months for the chemotherapy group. After 5 years of follow-up, 31.9% in the pembrolizumab group survived compared to 16,3% in the chemotherapy group (17).

It is important to note that KEYNOTE-024 excluded patients with Eastern Cooperative Oncology Group performance status (ECOG) higher than 1. The scale of ECOG status ranges from 0 to 5 and describes a patients functional status (18).

At grade 0 the patient can be fully active and carry out all tasks which they could perform prior to the disease, without restrictions (19).

At grade 1 the patient is restricted in carrying out physically strenuous activity, but can at times carry out light work such as cleaning or office work (19).

At grade 2 the patient is capable of self-care, but cannot work. The patient is active >50% of waking hours (19).

At grade 3 the patient is able to carry out limited self-care, and is confined to a bed or chair >50% of waking hours (19).

At grade 4 the patient is completely disabled, unable to carry out any self-care and bedridden (19).

At grade 5 the patient is dead (18).

1.5 Immature data

Pembrolizumab was approved based on immature data, which poses risks regarding the effectiveness of the drug, and the pharmacoeconomic assessment (2).

Immature data is quite prevalent and was used in 41% of oncology HTAs in the UK (20). In the UK managed access agreements (MAAs) are utilized when a drug is not recommended for routine use due to clinical uncertainty. Among the HTAs with MAAs, 87.5% were routinely commissioned after a median time of 36 months (21).

In Norway 35% of total reimbursement decisions were based on data with uncertainty regarding relative efficacy. Among the drugs with limited data, 47% were approved for reimbursement. Of the treatments approved despite uncertainty it was not advised by NOMA to calculate ICER for nearly half of them. In the context of oncology, 39% of therapies are uncertain, and 49% of drugs with uncertainty are approved versus 73% of drugs with more complete data |(22).

In the period after implementing the MAA approval of oncology HTAs increased from 59 to 72% and currently it is 78% (23). MAA may have contributed to increased access to new and innovative oncology treatments, but there may be other contributing factors. Compared to Norway around 66% of oncology HTAs are approved (22).

To increase access to cost-effective treatments implementing a system like the MAA may be beneficial for patients and the Norwegian healthcare system.

The publication of a five-year update on the KEYNOTE-024 provided us with the chance to compare initial data to long term findings in both real world patient and from the pivotal study.

1.6 Cancer registry of Norway

The dataset forming the foundation of this analysis is provided by CRN. Some of the data is manually reported with electronic forms, such as patient characteristics, but CRN utilizes multiple sources and methods of data capture for their registry. One of the most important sources of data for this thesis are IT- systems which are designed for drug-based cancer treatments. CRN also uses data retrieved from the Norwegian patient registry (NPR) to validate their own data. NPR collects data on treatments received in the specialist healthcare branch. The treatments are categorized by procedure codes and type of drug administered (24).

Unfortunately, these systems are not utilized in all healthcare regions, were implemented at different times, and we therefore lack data from multiple healthcare regions before 2019. Data on adjuvant treatment received on H-prescriptions is also unavailable until 2019. H-prescriptions are prescriptions financed by hospitals but dispensed at community pharmacies as a part of the treatment patients receive in the specialist healthcare (24).

Given the unavailability of H-prescription data prior to 2019, it's challenging to determine whether a patient has ALK or EGFR positive mutations, as this information is not included in the provided data. Post-2019, however, adjuvant treatments for patients with anaplastic lymphoma kinase (ALK) or epidermal growth factor receptor (EGFR) positive mutations are documented in H-prescriptions. The information on H-prescriptions is limited, and only provides information regarding dispensing date and type of drug. However the drugs used for adjuvant treatment of NSCLC has a limited number of approved indications.

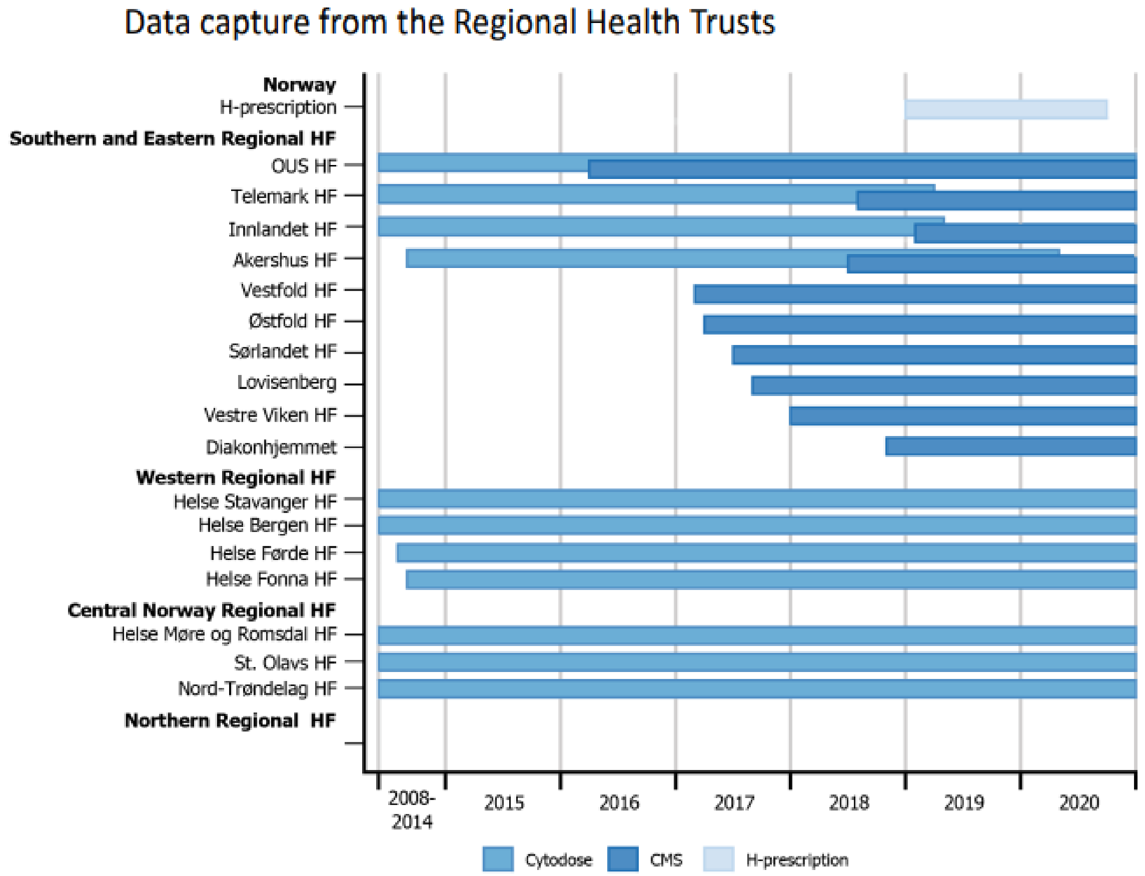


Figure 1.6.1: Overview of implementation and data capture from the regional health trusts (24).

1.7 Types of NSCLC

NSCLC accounts for 85% of all lung cancer and has three subtypes where adenocarcinoma is the most common. Adenocarcinoma accounts for about 40% of all lung cancer cases and is notably the predominant form of lung cancer among both non-smokers and smokers. It originates in the glandular epithelial cells and is commonly found in the outer regions of the lung. Adenocarcinomas tend to grow more slowly compared to other NSCLC types, which may allow for more effective treatment options if detected early. Smoking is the leading risk factor for lung cancer, accounting for at least 80% of lung cancer deaths. Other causes include radon exposure and air pollution (25).

Squamous-cell carcinoma constitutes 25-30% of all lung cancer diagnoses and predominantly originates from squamous cells in the epithelium of the bronchial tubes, situated centrally in the lungs. Large cell (undifferentiated) carcinoma, on the other hand, accounts for approximately 5-10% of lung cancer cases. This carcinoma is characterized by an absence of distinguishable squamous or glandular differentiation, frequently leading to its diagnosis via exclusionary criteria. The neoplasm commonly manifests in the central lung regions and has the propensity to metastasize to adjacent lymph nodes, the chest wall, and distant organs (25).

1.8 Treatment regimens

1.8.1 Pembrolizumab

Pembrolizumab is a monoclonal antibody that functions as an immune checkpoint inhibitor. It selectively targets the programmed cell death protein 1 (PD-1) receptor on T-cells. By binding to this receptor, pembrolizumab disrupts the interaction between PD-1 and its ligands, PD-L1 and PD-L2. This interaction is a key mechanism exploited by cancer cells to evade immune responses. Pembrolizumab is devoid of cytotoxicity (26).

Common adverse events include anemia (4.5%), pneumonia (1.9%), and pneumonitis (2.6%) based on the data from the KEYNOTE-024 study. Long-term side effects are immune-related illnesses such as diabetes and thyroiditis, although they were rare and inconsequential for the pharmacoeconomic model (2).

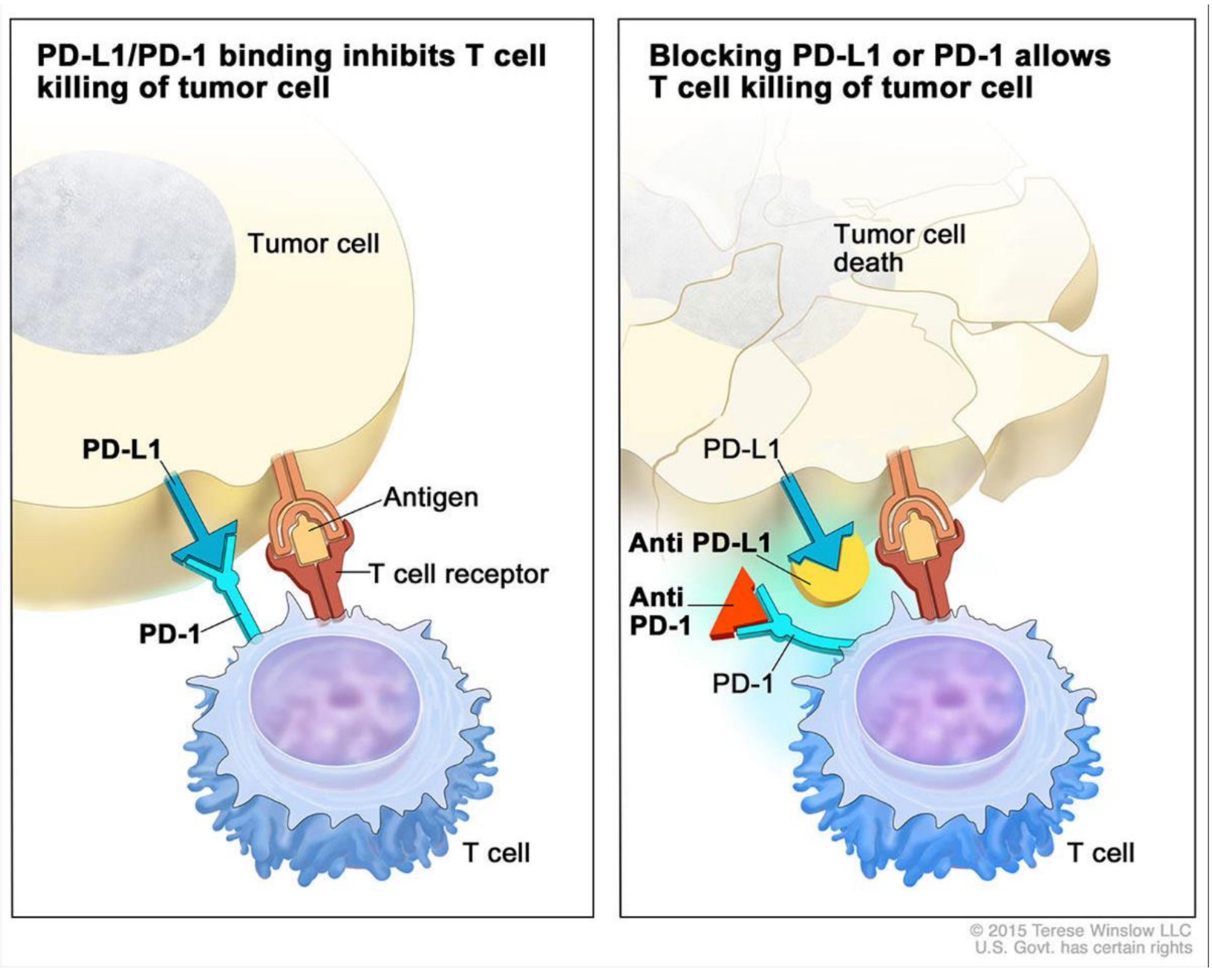

PD-L1/PD-1 binding inhibits T cell killing of tumor cell
Tumor cell
PD-L1
Antigen
T cell receptor
PD-1
T cell
Blocking PD-L1 or PD-1 allows T cell killing of tumor cell
Tumor cell death
PD-L1
Anti PD-L1
Anti PD-1
PD-1
T cell
© 2015 Terese Winslow LLC
U.S. Govt. has certain rights

Figure 1.8.1: Mechanism of action for pembrolizumab at the receptor site.

Due to the high cost of pembrolizumab, the margins for cost-effectiveness are small. Minor differences between efficacy and effectiveness can render the drug not cost-effective. Therefore the drug was first approved as a monotherapy for patients expressing PD-L1 over 50%. The estimated increased budget cost per patient is 48000 euro in 2017.According to a study conducted in northern Norway each patient should gain a mean life expectancy of nine months to make the treatment cost-effective as a second line therapy for NSCLC (27).

1.8.2 Previous standard of care

National guidelines in Norway recommend the use of carboplatin in combination with vinorelbine for the treatment of inoperable NSCLC in patients lacking ALK or EGFR mutations. This deviates from the comparator in the KEYNOTE studies, as pemetrexed is rarely used in clinical practice in Norway. According to NOMA’s assessment, the efficacy of different platinum-based chemotherapy regimens is comparable. Consequently, their pharmacoeconomic analysis is founded on the outcomes observed in the comparator arm of the KEYNOTE study, utilizing the medication costs associated with carboplatin and vinorelbine (2).

Carboplatin is one of the main platinum-based drugs. The target of carboplatin is DNA, and to interact with DNA, it must first cross the cell membrane and undergo hydrolysis, allowing it to form a covalent bond with the N7 position of purine bases. This inhibits transcription and replication and causes cell death. The mechanism of action is also responsible for the drug's cytotoxic effects (28).

Vinorelbine belongs to the class of vinca alkaloids. Its primary mechanism of action is to inhibit cell division by disrupting the assembly of microtubules. Disruption of the assembly of microtubules inhibits the mitotic spindle, which is essential for separating chromosomes during cell division. As a result, the cancer cells are unable to complete cell division, leading to cell cycle arrest and eventually cell death (29).

In the pharmacoeconomic assessment common side effects from platinum-based chemotherapy include anemia (23%), neutropenia (18%), pneumonia (7%) and thrombocytopenia (12%).Patients undergoing platinum-based chemotherapy exhibit a higher incidence of adverse reactions compared to those receiving pembrolizumab (2).

1.9 Changes in treatment guidelines for NSCLC during the study period

Decision Forum approved pembrolizumab as first-line treatment of locally advanced or metastatic PD-L1 positive NSCLC in patients without EGFR or ALK positive mutations and PD-L1 expression of at least 50% in May 2017.

During our follow-up period there were major changes in treatment guidelines regarding pembrolizumab and other immunotherapies for NSCLC. In April 2019, Decision Forum approved treatment with pembrolizumab combined with pemetrexed and platinum based chemotherapy in patients with adenocarcinoma histology without ALK,EGFR or proto-oncogene tyrosine-protein kinase-1 (ROS1) mutations and PD-L1 under 50%. Combination treatment is also approved for patients with PD-L1 above 50%, but according to current guidelines there is a high degree of uncertainty if there is an additive effect from chemotherapy above PD-L1 75% (30).

Combination therapy with pembrolizumab, paclitaxel and carboplatin for squamous cell carcinoma was approved in October 2020 in patients with PD-L1 expression lower than 50%. The current guidelines also recommend combination therapy for patients with PD-L1 expression under 50% unless chemotherapy is not tolerated, and there is also a high degree of uncertainty regarding additive effect with expression above 75%.

The alterations in these guidelines may lead to a decrease in the cohort of pembrolizumab monotherapy patients within our study period. Additionally, this might result in monotherapy patients showing an elevated frequency of PD-L1 expression beyond 75%, possibly causing the effect of the drug to be overestimated. The patients enrolled after these guidelines changes also have a shorter follow-up.

1.10 Pharmacoeconomic assessments abroad

Studies conducted in other countries have shown that the treatment of NSCLC with pembrolizumab is not cost-effective, but the economic threshold for implementation was lower than the Norwegian threshold. In addition, the main unit for evaluating effectiveness were life years gained rather than QALYs gained (31). A systematic review has shown that the current cost-effectiveness studies are of moderate quality, and the decision analytic modeling methods have potential for improvement. The methodology for pharmacoeconomic research varies between countries, and studies concluded pembrolizumab was cost-effective in the United States and Switzerland, but not China, France, the UK or Singapore (32).

Varying effectiveness, willingness to pay, pharmacoeconomic methods, medical guidelines and treatment expenses across nations contribute to a limited alignment between study outcomes. The study outcomes are neither easily transferable to a cost-effectiveness analysis done by the Norwegian government, since the willingness to pay varies based on the severity of the disease (1). Therefore, a study on a Norwegian population is beneficial, with the same expenses and pharmacoeconomic considerations.

1.11 Ethical considerations

In this thesis, we will handle large amounts of patient data. The risk of sensitive data leakage is significant and could affect thousands. Therefore, all patient data are anonymized, ensuring stringent adherence to data protection requirements from the Directorate of e-health. Furthermore, the research aligns with the University of Tromsø’s ethical guidelines.

Since there is a large amount of data and variables, it might be possible to link anonymous data to individuals based on the rarity of combining variables such as height, sex, diagnosis, and age.

The data that’s not stored locally is encrypted with Azure Information Protection. This ensures that in the event of a security breach, the data at risk remains secure and inaccessible to unauthorized individuals.

2 Research question

Can real world data be effectively utilized for pharmacoeconomic assessments in addition to pivotal controlled trials?

2.1 Aims

This thesis aims to re-evaluate the cost-effectiveness of pembrolizumab in treating NSCLC within the Norwegian healthcare system with real world data.

2.2 Objectives

1. To re-evaluate the economic feasibility of pembrolizumab in treating lung cancer within the Norwegian healthcare framework.

2. To identify the possible variables contributing to the gap between its efficacy in controlled environments and its effectiveness in a clinical setting.

3. Provide a framework for assessing treatments with immature data.

3 Methodology

The clinical data spans between 2016-2022 and there are 3592 patients included in the initial dataset. We will use publicly available pricing data for pembrolizumab, due to the confidential nature of pricing agreements in Norway.

Analytical techniques include Cox regression and Kaplan Meier survival analysis. Cox regression will be used to estimate which variables are associated with changes in our outcomes such as QALYs, overall survival, and progression-free survival. Kaplan-Meier will be used to compare overall survival between our data and KEYNOTE-024. Level of significance for all relevant methods is $p < 0.05$.

The data includes diagnosis, time of diagnosis, start and end dates for radiation treatment, and, if applicable, status date (emigrated/deceased). The data also encompasses treatment information, including medications, functional level, and individual characteristics such as age and sex. For more information see table 2.

Sensitivity analysis will be conducted to examine factors that may influence on our results. We will investigate differences in survival between patient pre and post 2019 due to missing information regarding adjuvant treatment and due to a higher degree of uncertainty regarding the validity of patient characteristic variables. We will also conduct a sensitivity analysis pre and post implementation of pembrolizumab as first line treatment, since the treatment guidelines are updated during our follow-up.

Portions of the text and code in this thesis were refined and debugged with the aid of ChatGPT-4, aiming to enhance the readability and quality of the text, as well as the functionality of the code. Recommendations from ChatGPT were adopted and adjusted based on the clarity of the text or the robustness of the code.

More specifically ChatGPT-4 was fed code chunks with errors or lines of text, prompted with sentences such as “suggest a more concise text” or “what is the source of this error”.

3.1 Kaplan-Meier

The Kaplan-Meier method is a statistical technique used to estimate the survival function from lifetime data. It is used to measure the fraction of patients living for a certain amount of time after treatment or diagnosis. The Kaplan-Meier curve is a graphical representation of this survival function, offering visualization of the probability of survival over time.

This curve allows assessment of the survival rates between groups of patients, such as those receiving pembrolizumab compared to standard treatment. It is a useful tool for understanding overall treatment effectiveness, and how it varies over time, since it provides a straightforward graphical representation.

While using the Kaplan-Meier method, several analytical considerations are vital to ensure valid and reliable results. Censoring should be non-informative, meaning that the reasons for censoring should be independent of the probability of the event of interest. Log-rank test will be used to determine if the difference between the groups is significant. Survival time will also be reported with confidence intervals.

We will compare our Kaplan-Meier curves with clinical studies to visually highlight differences in overall survival.

3.2 Cox regression

Cox regression, also known as the Cox Proportional-Hazards model, is a statistical method used in the analysis of survival data. It is an extension of the Kaplan-Meier method and allows for the inclusion of additional variables that might affect the outcome, instead of only a single categorical variable. The Cox model estimates the impact of multiple variables and how they influence the probability of an event happening, such as death, over time.

This may help to identify which variables are associated with the efficacy – effectiveness gap.

While using the Cox regression, several analytical considerations are vital to ensure valid and reliable results. Linearity, interaction, confounding and multicollinearity between variables must be considered. The proportional hazards assumption should be evaluated with methods such as the Schoenfield residual test. The observations should also be independent.

If the necessary conditions are met, a model is chosen based on the Akaike information criterion (AIC). The AIC helps to choose the best model by balancing how well the model fits the data against the complexity of the model.

The data on covariates are limited and only measured at time of diagnosis, and are likely to vary at various times due to disease progression, mutations and patient health. Consequently, it is probable that the assumption of proportional hazards may be compromised. Nevertheless, these covariates can still serve as valuable indicators of survival at the time of diagnosis.

However, their applicability in predicting future outcomes diminishes in cases of patients with extensive follow-up.

3.3 Treatment of raw data

Our data was delivered in 4 separate datasets from CRN. Some of the variables in the dataset often overlap and provide the same information, such as different codes for the same diagnosis. Each patient was identified with a unique patient ID (PID), and a unique illness ID (SID) in all the datasets. All dates in the datasets were set to the 15th every month by CRN, to avoid identifying patients based on date diagnosis or treatments.

The first dataset contained patient variables such as ECOG status, age, sex and multiple variables describing the tumor location and morphology. This dataset had missing data, some of which were critical for our analysis such as ECOG status, cancer stage and PD-L1 levels.

The second dataset contained information about prescribed treatments received from the hospital. Variables such as active ingredient, ATC-code and treatment regimen, and type of treatment. Treatment date is also included, but since the time variable is the 15th every month they are sometimes registered as a single administration.

The treatment regimen is determined by the cancer drugs administered to the patient. E.g. using carboplatin and vinorelbine constitutes the treatment regimen, and the treatment type is chemotherapy. Use of antiemetics falls under the carboplatin and vinorelbine regimen, but the treatment type is supportive care.

The third dataset contains data from H-prescriptions, capturing information on the drug and the dispensing date.

The final dataset contains information regarding radiation therapy with variables such as intention for radiation, dosage and date.

3.4 Survival time

The duration of survival in our study is measured from the initial administration of pembrolizumab monotherapy until death or censoring. Therefore, in cases where a patient undergoes combination therapy up to three months before transitioning to monotherapy, the survival period starts with the first administration of pembrolizumab monotherapy, instead of the onset of combination therapy.

We will use restricted mean survival time (RMST) as a measure of overall survival. The restricted mean survival time will be calculated based on follow-up period, and a separate RMST will be extrapolated to 20 years follow-up using the survextrap package (33).

Certain cases of NSCLC were excluded due to patients having multiple cases of NSCLC. These patients had cases several years prior to the introduction of pembrolizumab, some of which were before 2010. The exclusion of these cases are anticipated to have no impact on survival analysis, as survival duration is calculated from the initiation of pembrolizumab monotherapy treatment.

3.5 Types of missingness

In our dataset, a sizable portion of data is missing, predominantly in the areas of ECOG-status, cancer stage, and PD-L1 values. This issue predominantly stems from the transition to a new reporting system, though it's important to note that missing values are probably not related to clinical outcomes. The term 'not reported' is used when specific values are left blank in submissions, while 'missing' refers to instances where the entire form was not submitted. In several cases, values were either unknown or not explicitly specified.

Given the crucial nature of these variables in relation to our study's endpoint, we will impute these missing values. Afterwards we will conduct a comparative analysis between this imputed dataset and a dataset where the missing values are left blank.

3.6 Imputation

Missing data was imputed with the R-package Mice. The imputation was stepwise in the following order: cancer stage, ECOG score, morphology group, and PDL1 results. Ten imputations were conducted with ten iterations, and with 123456 as a seed code for the imputation.

3.7 Extrapolation

Extrapolations were conducted with spline models using the survextrap package (33). We calculated RMST with the same 20-year timeframe used in the assessment by NOMA (2).

3.8 Digitizer

We approached MSD to inquire if access to their survival data from the KEYNOTE studies was possible. Due to time constraints, our only alternative was to digitize the published

Kaplan-Meier curves of KEYNOTE-024. We did receive the data provided to NOMA during the assessment, but not long-term data from the study.

The digitization process involved NOMA's estimation of Overall Survival (OS) and the 5-year update from KEYNOTE-024. (2, 34).

The digitization of the figures was conducted by saving them using a snipping tool, followed by uploading to WebPlotDigitizer. In this tool, the X and Y axes were defined, and the curve was carefully annotated with points to ensure accuracy. The data points acquired were then uploaded to the Enhanced Kaplan-Meier Curves shiny app, including intervals for the number of patients at risk. Finally, the individual patient data downloaded from this process was used for survival analysis in R
(35,36).

3.9 Inclusion criteria

We created two patient groups.

Group 1 are patients that received pembrolizumab monotherapy, regardless of clinical status at the time of diagnosis. Since the measurement of ECOG status and other clinical markers are only recorded once, we assumed that patients would receive monotherapy if they fulfilled the requirements. This group will represent real world data. Please see the consort diagram for information regarding selection criteria.

Group 2 has been designed with selection criteria to closely mirror the conditions of a clinical trial. This cohort includes only those patients who were diagnosed in stage IV, exhibited a PD-L1 level ≥ 50%, having an ECOG performance status ≤ 1 or lower, and had not received any adjuvant treatments. Given the prevalence of incomplete data for these variables among our patient population, we have imputed missing values.

We also conducted a sensitivity analysis by removing patients that died within 1 months of pembrolizumab initiation, to simulate exclusion of patients that were excluded due to survival prognosis under 3 months.

3.9.1 Patient groups in our analysis

If the patient groups show no significant differences with a Kaplan-Meier curve and Log-Rank test, Group 1 will serve as the reference for conducting pharmacoeconomic analyses and to compare with clinical studies. The rationale for choosing Group 1 stems from a closer representation of real world clinical practice, and the dataset contains more observations, reducing the uncertainty in our estimations.

3.10 Pharmaeconomic assessments

Unfortunately due to data limitations we can not use QALYs in our pharmacoeconomic assessment. We have no data on adverse events or disease progression, so we can only use lifespan. NOMAs analysis is heavily influenced by changes regarding disease progression, since it both affects QALYs, and treatment cost.

To calculate total cost we used the estimated treatment costs mentioned in the pharmacoeconomic assessment, combined with real world data regarding doses administered. We calculated the average doses received and multiplied dosages received with administration and drug cost.

We will also multiply average survival time with either the weekly cost of progression free survival or survival with progression, which amounts to 657 and 1824 NOK respectively. This will provide a range of the possible ICER value. Finally we will estimate the cost of death, which is 50382 NOK. This cost will be calculated by number of events divided by total patients (2).

Due to the uncertainty of rounded dates, we will also estimate a worst-case scenario with one month reduced survival and with disease progression for the entire follow-up duration.

This method differs from the method utilized by NOMA. Due to several limitations which makes it impossible to accurately calculate ICER according to NOMAs methods, we opted to utilize a simplified cost calculation in our analysis.

4 Results

The dataset contain 3642 cases of lung cancer and 3593 patients. After data cleaning and inclusion criteria 1347 cases and patients remain. For an overview of patient characteristics see table 1 and 2. Our median follow-up duration was 8 months due to the continuous inclusion period, maximum follow-up was 67 months until the cut-off after 31th December 2022.

It is important to be aware of different characteristics in the real world group compared to KEYNOTE-024 before interpreting the results. Unlike KEYNOTE-024, 29% of our patients has a ECOG score of 2 or higher while KEYNOTE-024 excluded patients with ECOG status ≥ 2. 30% of the cases in the real world group were squamous cell carcinomas, compared to 19% in KEYNOTE-024. Median age was also higher, 71 years in the real world group compared to 65 in KEYNOTE-024 (34).

A higher proportion of squamous cell carcinomas may contribute to increased mortality, both due to comorbidities associated with the histology, and the histology itself (37,38).

4.1 Groups

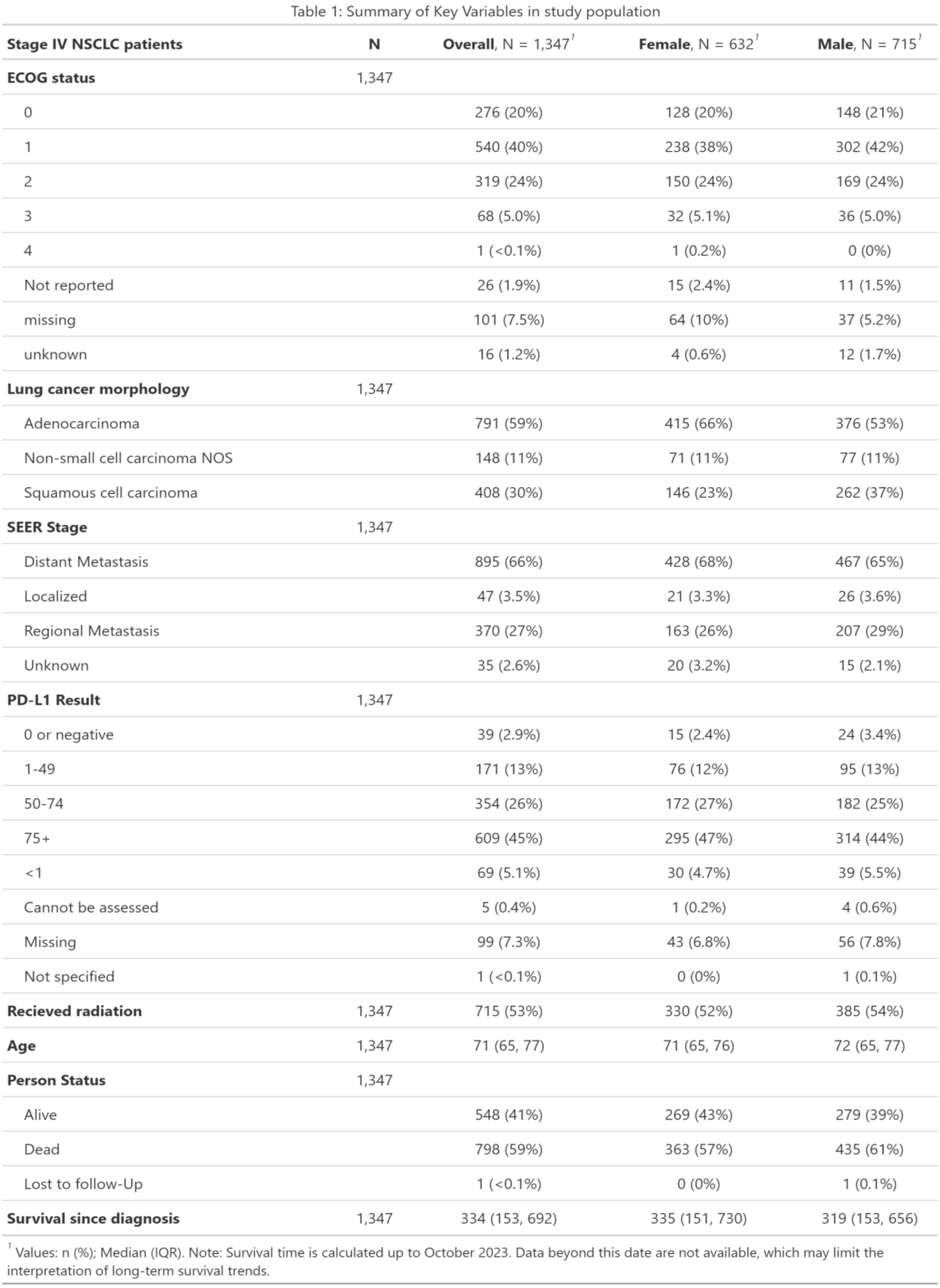


Table 1: Summary of Key Variables in study population

Stage IV NSCLC patients	N	**Overall**, N = 1,347[1]	**Female**, N = 632[1]	**Male**, N = 715[1]
ECOG status	1,347			
0		276 (20%)	128 (20%)	148 (21%)
1		540 (40%)	238 (38%)	302 (42%)
2		319 (24%)	150 (24%)	169 (24%)
3		68 (5.0%)	32 (5.1%)	36 (5.0%)
4		1 (<0.1%)	1 (0.2%)	0 (0%)
Not reported		26 (1.9%)	15 (2.4%)	11 (1.5%)
missing		101 (7.5%)	64 (10%)	37 (5.2%)
unknown		16 (1.2%)	4 (0.6%)	12 (1.7%)
Lung cancer morphology	1,347			
Adenocarcinoma		791 (59%)	415 (66%)	376 (53%)
Non-small cell carcinoma NOS		148 (11%)	71 (11%)	77 (11%)
Squamous cell carcinoma		408 (30%)	146 (23%)	262 (37%)
SEER Stage	1,347			
Distant Metastasis		895 (66%)	428 (68%)	467 (65%)
Localized		47 (3.5%)	21 (3.3%)	26 (3.6%)
Regional Metastasis		370 (27%)	163 (26%)	207 (29%)
Unknown		35 (2.6%)	20 (3.2%)	15 (2.1%)
PD-L1 Result	1,347			
0 or negative		39 (2.9%)	15 (2.4%)	24 (3.4%)
1-49		171 (13%)	76 (12%)	95 (13%)
50-74		354 (26%)	172 (27%)	182 (25%)
75+		609 (45%)	295 (47%)	314 (44%)
<1		69 (5.1%)	30 (4.7%)	39 (5.5%)
Cannot be assessed		5 (0.4%)	1 (0.2%)	4 (0.6%)
Missing		99 (7.3%)	43 (6.8%)	56 (7.8%)
Not specified		1 (<0.1%)	0 (0%)	1 (0.1%)
Recieved radiation	1,347	715 (53%)	330 (52%)	385 (54%)
Age	1,347	71 (65, 77)	71 (65, 76)	72 (65, 77)
Person Status	1,347			
Alive		548 (41%)	269 (43%)	279 (39%)
Dead		798 (59%)	363 (57%)	435 (61%)
Lost to follow-Up		1 (<0.1%)	0 (0%)	1 (0.1%)
Survival since diagnosis	1,347	334 (153, 692)	335 (151, 730)	319 (153, 656)

[1] Values: n (%); Median (IQR). Note: Survival time is calculated up to October 2023. Data beyond this date are not available, which may limit the interpretation of long-term survival trends.

Table 1: Patient characteristics in the real world group categorized by gender.

Table 2: Summary of Key Variables in the Whole Group

Overview of all cases by cancer type	**N**	**Overall**, N = 3,642[1]	**I**, N = 131[1]	**II**, N = 143[1]	**III**, N = 833[1]	**IV**, N = 2,178[1]	**missing**, N = 305[1]	**unknown**, N = 52[1]
ECOG status	3,642							
0		993 (27%)	51 (39%)	57 (40%)	289 (35%)	582 (27%)	0 (0%)	14 (27%)
1		1,479 (41%)	62 (47%)	53 (37%)	373 (45%)	970 (45%)	0 (0%)	21 (40%)
2		629 (17%)	16 (12%)	29 (20%)	132 (16%)	442 (20%)	0 (0%)	10 (19%)
3		119 (3.3%)	1 (0.8%)	4 (2.8%)	23 (2.8%)	88 (4.0%)	0 (0%)	3 (5.8%)
4		4 (0.1%)	0 (0%)	0 (0%)	0 (0%)	4 (0.2%)	0 (0%)	0 (0%)
Not reported		64 (1.8%)	0 (0%)	0 (0%)	7 (0.8%)	57 (2.6%)	0 (0%)	0 (0%)
missing		305 (8.4%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	305 (100%)	0 (0%)
unknown		49 (1.3%)	1 (0.8%)	0 (0%)	9 (1.1%)	35 (1.6%)	0 (0%)	4 (7.7%)
Lung cancer morphology	3,642							
Adenocarcinoma		2,293 (63%)	80 (61%)	61 (43%)	391 (47%)	1,538 (71%)	199 (65%)	24 (46%)
Large cell carcinoma		3 (<0.1%)	1 (0.8%)	0 (0%)	1 (0.1%)	1 (<0.1%)	0 (0%)	0 (0%)
Large cell neuroendocrine carcinoma		27 (0.7%)	1 (0.8%)	3 (2.1%)	2 (0.2%)	17 (0.8%)	4 (1.3%)	0 (0%)
Non-small cell carcinoma NOS		401 (11%)	7 (5.3%)	18 (13%)	82 (9.8%)	252 (12%)	30 (9.8%)	12 (23%)
Other		2 (<0.1%)	2 (1.5%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Small cell carcinoma		6 (0.2%)	0 (0%)	1 (0.7%)	3 (0.4%)	2 (<0.1%)	0 (0%)	0 (0%)
Squamous cell carcinoma		907 (25%)	38 (29%)	60 (42%)	354 (42%)	367 (17%)	72 (24%)	16 (31%)
missing		3 (<0.1%)	2 (1.5%)	0 (0%)	0 (0%)	1 (<0.1%)	0 (0%)	0 (0%)
SEER Stage	3,642							
Localized		134 (3.7%)	83 (63%)	28 (20%)	13 (1.6%)	0 (0%)	6 (2.0%)	4 (7.7%)
Regional Metastasis		1,049 (29%)	46 (35%)	101 (71%)	784 (94%)	5 (0.2%)	76 (25%)	37 (71%)
Distant Metastasis		2,331 (64%)	2 (1.5%)	14 (9.8%)	34 (4.1%)	2,171 (100%)	106 (35%)	4 (7.7%)
Unknown		128 (3.5%)	0 (0%)	0 (0%)	2 (0.2%)	2 (<0.1%)	117 (38%)	7 (13%)
PD-L1 precentage	3,642							
0 or negative		228 (6.3%)	5 (3.8%)	2 (1.4%)	40 (4.8%)	166 (7.6%)	15 (4.9%)	0 (0%)
1-49		904 (25%)	28 (21%)	34 (24%)	242 (29%)	498 (23%)	87 (29%)	15 (29%)
50-74		618 (17%)	29 (22%)	31 (22%)	140 (17%)	370 (17%)	42 (14%)	6 (12%)
75+		908 (25%)	29 (22%)	48 (34%)	214 (26%)	521 (24%)	76 (25%)	20 (38%)
<1		460 (13%)	20 (15%)	12 (8.4%)	83 (10.0%)	305 (14%)	36 (12%)	4 (7.7%)
Cannot be assessed		15 (0.4%)	0 (0%)	0 (0%)	4 (0.5%)	11 (0.5%)	0 (0%)	0 (0%)
Missing		508 (14%)	20 (15%)	16 (11%)	110 (13%)	306 (14%)	49 (16%)	7 (13%)
Not specified		1 (<0.1%)	0 (0%)	0 (0%)	0 (0%)	1 (<0.1%)	0 (0%)	0 (0%)
Age (years)	3,642	70 (63, 75)	72 (67, 76)	72 (66, 77)	70 (64, 76)	69 (62, 74)	69 (63, 74)	72 (66, 76)
Person Status	3,642							
Alive		1,509 (41%)	64 (49%)	74 (52%)	444 (53%)	797 (37%)	109 (36%)	21 (40%)
Dead		2,132 (59%)	67 (51%)	69 (48%)	389 (47%)	1,380 (63%)	196 (64%)	31 (60%)
Lost to follow-Up		1 (<0.1%)	0 (0%)	0 (0%)	0 (0%)	1 (<0.1%)	0 (0%)	0 (0%)
Survival since diagnosis (days)	3,642	457 (243, 822)	760 (395, 1,096)	607 (395, 958)	607 (395, 944)	395 (184, 699)	485 (243, 1,034)	549 (388, 951)
Gender	3,642							
Female		1,692 (46%)	54 (41%)	65 (45%)	354 (42%)	1,030 (47%)	168 (55%)	21 (40%)
Male		1,950 (54%)	77 (59%)	78 (55%)	479 (58%)	1,148 (53%)	137 (45%)	31 (60%)
Radiation therapy recieved	3,642	2,147 (59%)	80 (61%)	87 (61%)	576 (69%)	1,183 (54%)	189 (62%)	32 (62%)

[1] Values: n (%); Median (IQR). Note: Survival time is calculated up to October 2023. Data beyond this date are not available, which may limit the interpretation of long-term survival trends.

Table 2: Lung cancer cases in the dataset, categorized by cancer stage. Note that some patients have had more than one instance of cancer. This dataset includes all patients diagnosed with NSCLC that have received immunotherapy treatment.

The term “unknown” is used when CRN receives an electronic form where the clinician has marked the status as unknown. “Not specified” is indicated when CRN receives a form with no information filled in.

There was no significant difference in survival between the non-imputed real-world group (Group 1) and the imputed more selective group emulating the clinical trial (Group 2). Therefore, all the results presented are based on Group 1.

4.2 Overall survival in real world patients

Figure 4.2.1 displays the Kaplan-Meier curve of group 1. The median OS was 13 months (95% CI 12 - 15), with a mean OS of 25.6 months. Some of these patients did not remain on monotherapy and switched to chemotherapy or combination therapy. Overall survival after 5 years was 22.45%.

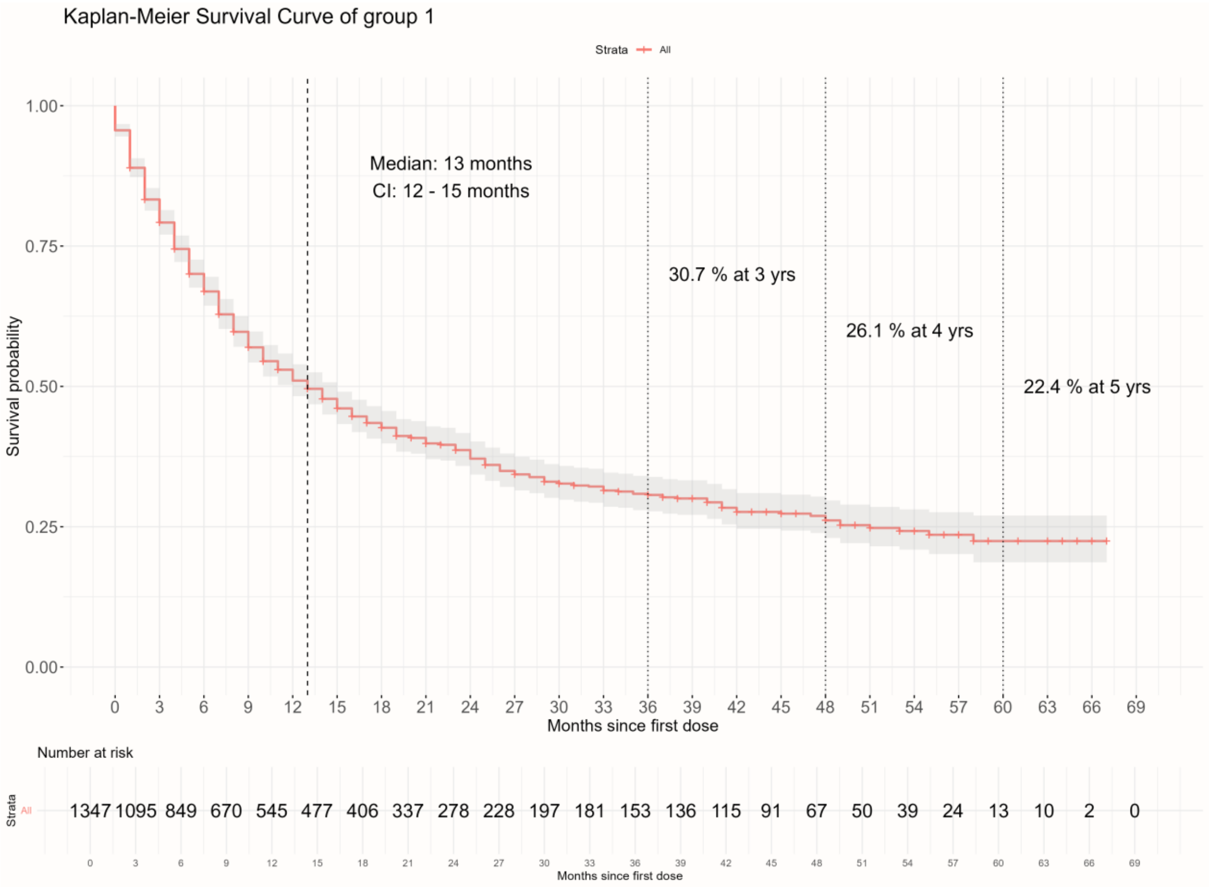


Figure 4.2.1: Kaplan-Meier survival curve of our real world patients in Group 1.

4.2.1 Real world overall survival compared to KEYNOTE-024

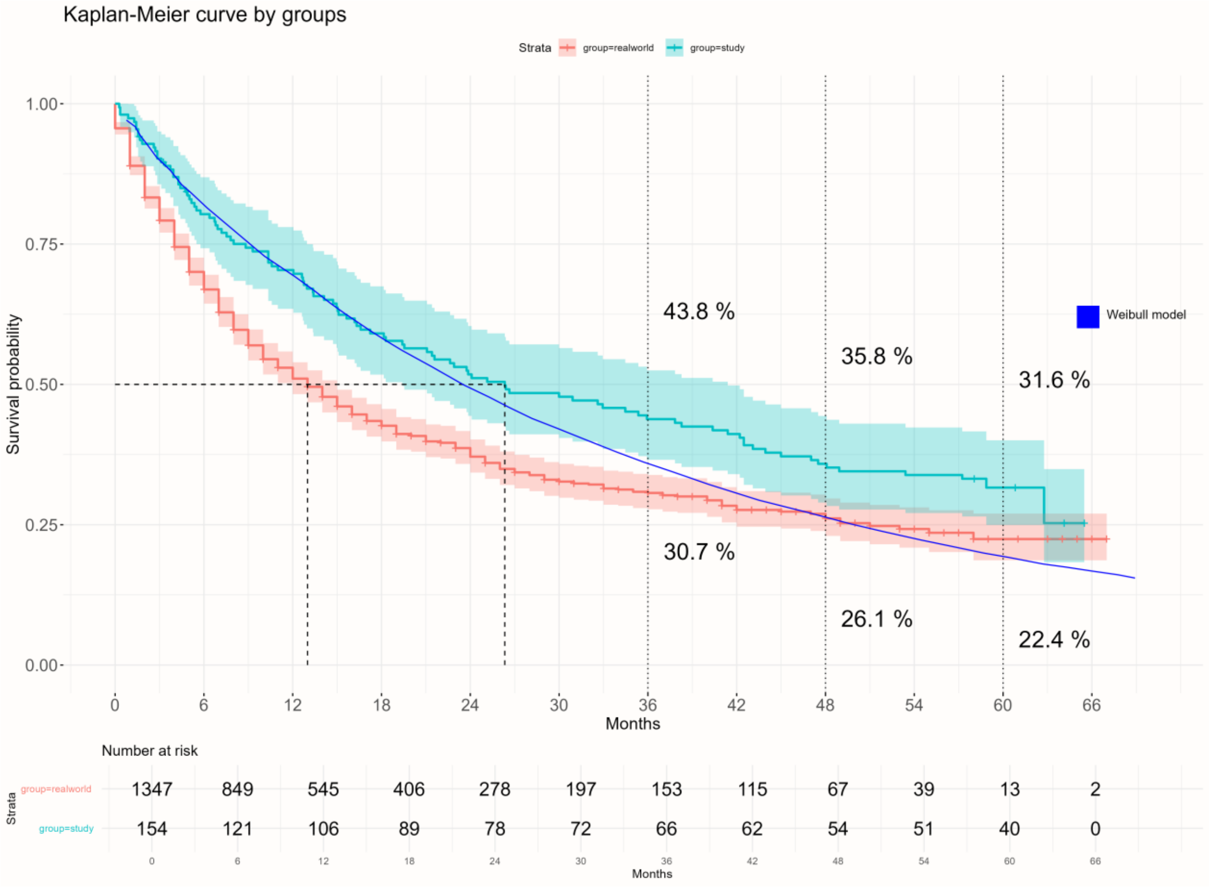


Figure 4.2.2: Kaplan-Meier survival curve of our real world patients in Group 1 compared to NOMAs Weibull extrapolation and KEYNOTE-024 patients.

Figure 4.2.2 displays the real world Kaplan-Meier compared to the 5-year update of the KEYNOTE-024 study and NOMAs Weibull distribution based on preliminary data from KEYNOTE-024. There is a significant gap in survival between real-world data and the the digitized survival data from KEYNOTE-024 (p= 0.001). Extrapolated survival based on the Weibull yielded an average life years gained of 2.72 according to NOMA (2).

Our digitized version of the 5-year update on KEYNOTE-024 seems to match the values in the original paper, see table 3. The mean OS was 33.6 months in the digitized KEYNOTE-024 group.

	KEYNOTE-024	Digitized version
Median OS in months (95% CI)	26.3 (18.3 - 40.4)	26.3 (19.4 - 41.4)
OS after 48 months (%)	38.8	38.9

Table 3: This table displays the median survival and survival after 48 months in the KEYNOTE-024 pembrolizumab group, compared to our digitized version.

4.2.2 Emulating the inclusion criteria

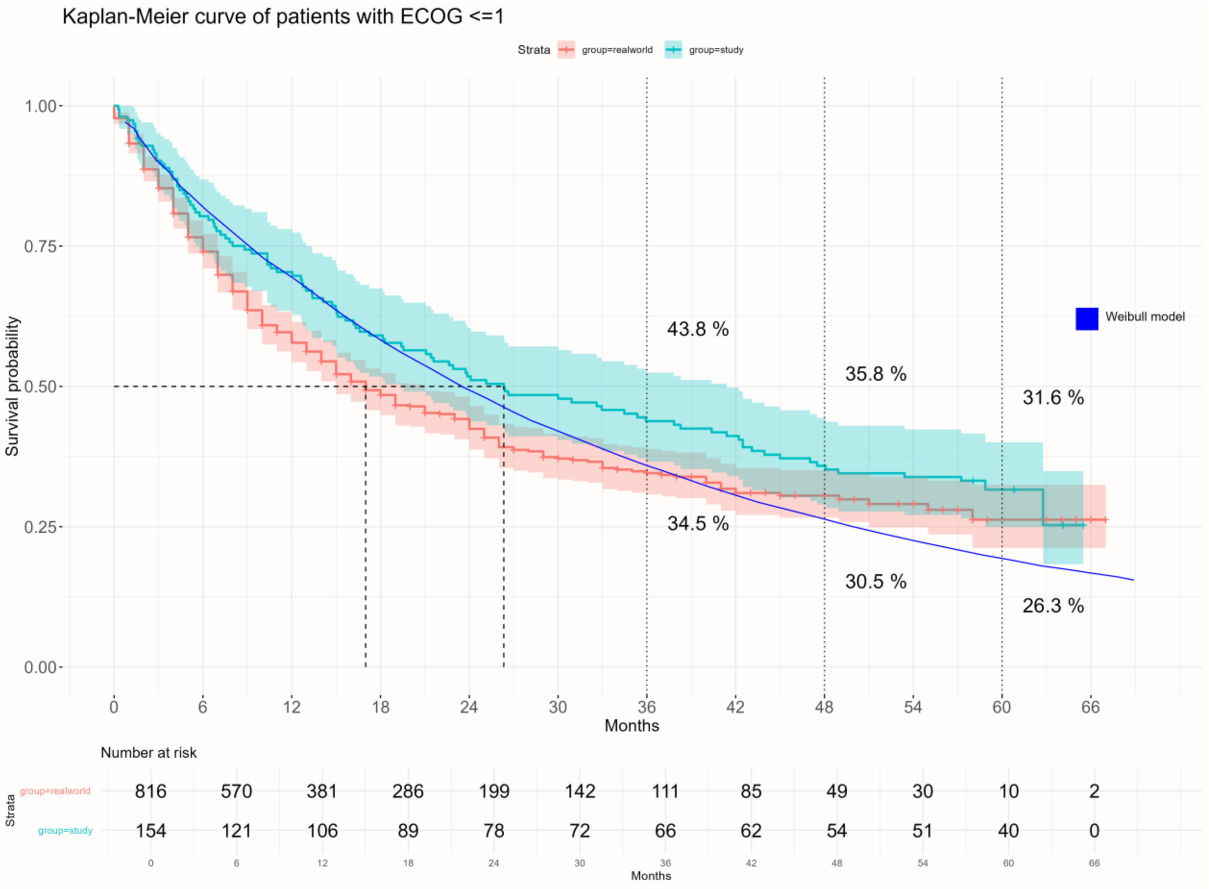


Figure 4.2.3: This Kaplan-Meier curve illustrates the overall survival of patients who had an ECOG level of ≤ 1 at diagnosis. Patients with unknown ECOG level were excluded.

To examine factors that may be associated with the efficacy - effectiveness gap we created a subgroup of patients with an ECOG status of <=1 within the real world group. A Log-Rank test was conducted, and the subgroup is not significantly different (p = 0.07) from the KEYNOTE-024 patients, and long-term survival seems to converge.

This exclusion of patients with an ECOG status higher than 1 ensures a cohort more closely aligned with the population of KEYNOTE-024, although it does not represent real world patients.

It’s crucial to re-emphasize that ECOG status is only recorded at time of diagnosis, 66% of patients are diagnosed at stage IV.

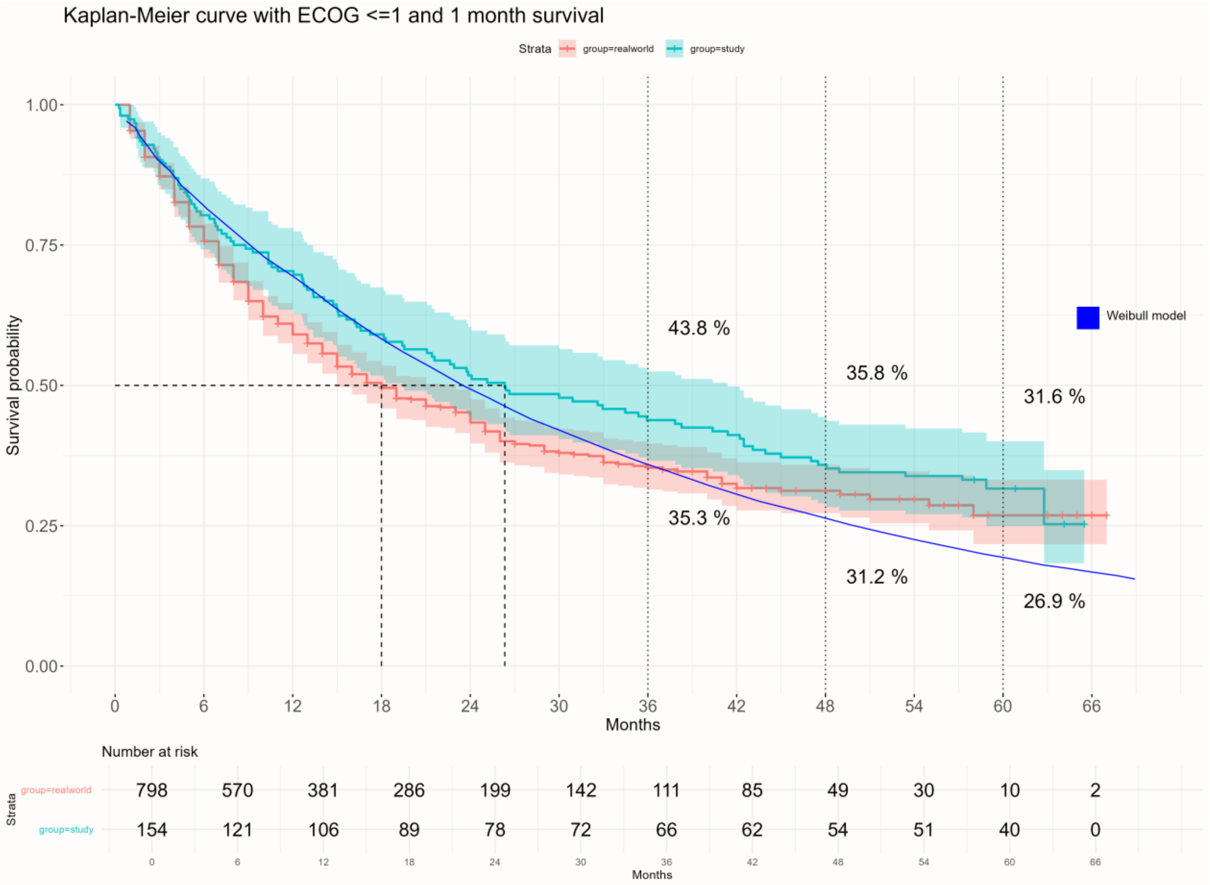


Figure 4.2.4: This Kaplan-Meier curve is based on our real-world data, with the exclusion of patients who passed away within the first 30 days and had a ECOG level of 0 or 1.

The method utilized in figure 4.2.4 is not methodologically sound, as it incorporates immortal time bias. Our intention was to emulate the KEYNOTE study’s criteria, which requires a minimum prognosis of three months and ECOG 0 or 1, by excluding patients who succumbed within 30 days of commencing treatment. For mean survival see table 7.

4.2.3 Patients with poor ECOG status

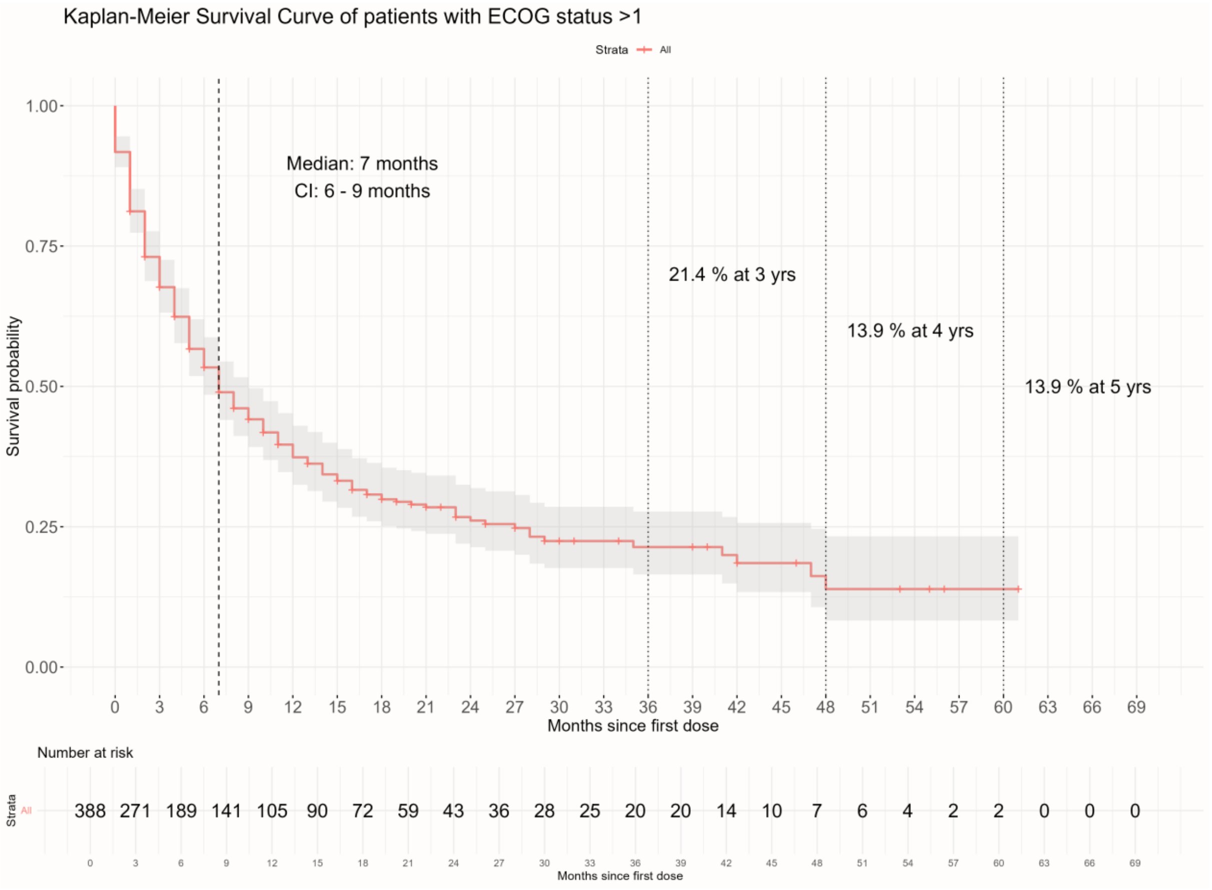


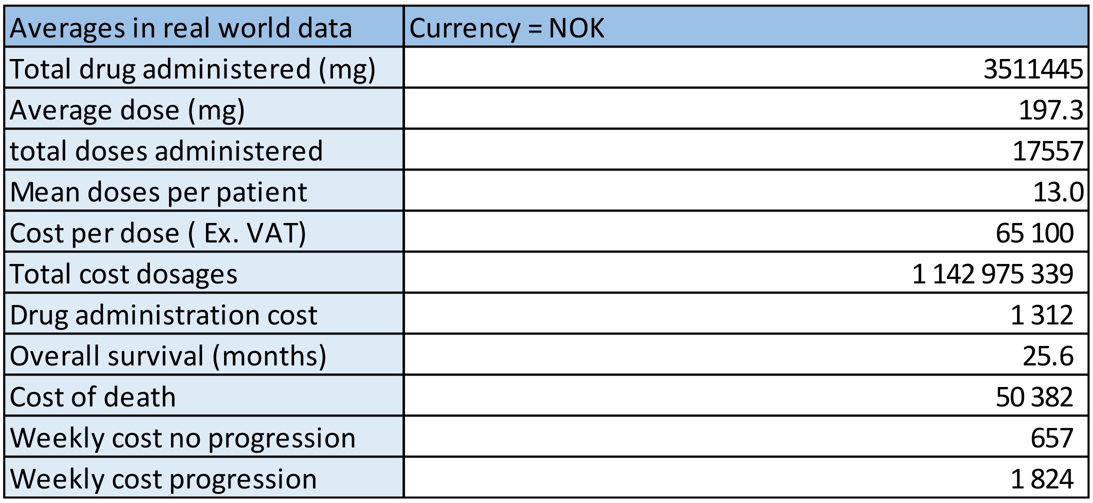
Figure 4.2.5: This Kaplan-Meier curve illustrates the overall survival of patients who had an ECOG level of ≥ 2 at the time of diagnosis. Patients with unknown ECOG level were excluded.

The median survival for real world patients that have an ECOG of ≥ 2 is 7 months as shown in figure 4.2.5. The mean age of this subgroup was 71.9 years, and 33.5% of the patients were diagnosed with squamous cell carcinoma. 17.5% had a ECOG status of three, and a single patient with ECOG status of four. This subgroup were most the patients removed from figure 4.2.3 and 4.2.4.

4.3 Pharmacoeconomic analysis

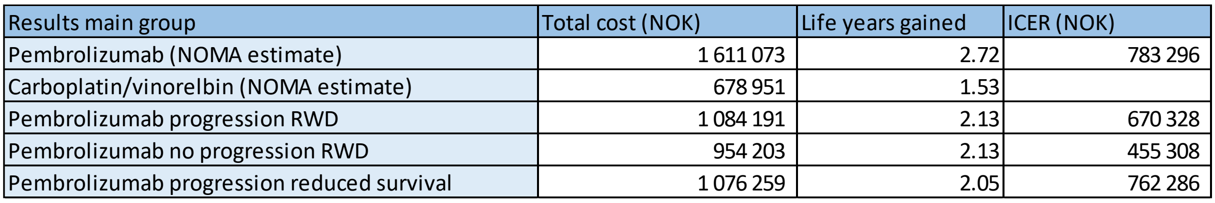
It is important to note that the analysis is based on average survival in the real world population, with no extrapolations. Prices or survival has also not been discounted and are based on unit prices from the HTA (2).

Unfortunately, due to the lack of a real world chemotherapy group we used NOMAs estimations for chemotherapy costs, which were based on the KEYNOTE-024 chemotherapy group (2,17). These groups have different characteristics, so they are not truly comparable.



Averages in real world data	Currency = NOK
Total drug administered (mg)	3511445
Average dose (mg)	197.3
total doses administered	17557
Mean doses per patient	13.0
Cost per dose (Ex. VAT)	65 100
Total cost dosages	1 142 975 339
Drug administration cost	1 312
Overall survival (months)	25.6
Cost of death	50 382
Weekly cost no progression	657
Weekly cost progression	1 824

Table 4: Example of the basis for cost calculations. The unit prices are constant, while mean doses and mean overall survival are calculated for every group. These results are then further multiplied with the unit prices above.

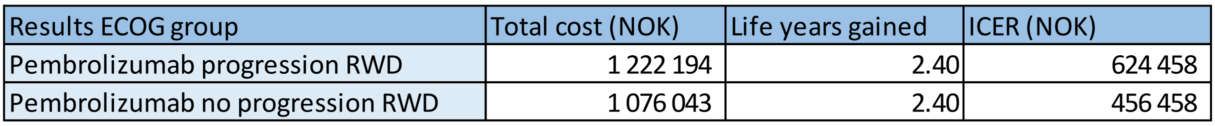


Results main group	Total cost (NOK)	Life years gained	ICER (NOK)
Pembrolizumab (NOMA estimate)	1 611 073	2.72	783 296
Carboplatin/vinorelbin (NOMA estimate)	678 951	1.53	
Pembrolizumab progression RWD	1 084 191	2.13	670 328
Pembrolizumab no progression RWD	954 203	2.13	455 308
Pembrolizumab progression reduced survival	1 076 259	2.05	762 286

Table 5: ICER of real world patients. The first two rows displays NOMAs estimates.

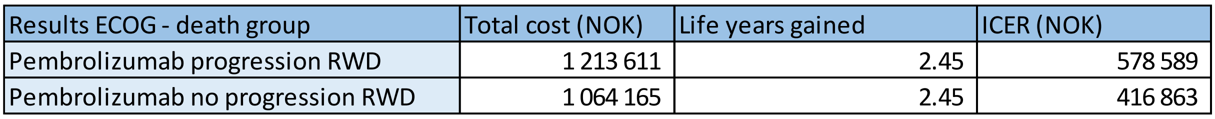
Table 5 displays the ICER for pembrolizumab compared to SoC. The values of the SoC are directly based off NOMAs main analysis. The next two rows contain separate calculations, with either progression during the entire follow-up period, or no progression at all as two extreme cases which gives a range estimate for ICER. The longest duration of follow-up for a patient was 67 months, until the cut-off point of 2023.

The final row depicting progression and decreased survival serves as a sensitivity analysis concerning the temporal variable. Recognizing that survival time may vary by approximately ±30 days, we decided to reduce the mean survival time to estimate a worst-case scenario, accounting for progression throughout the entire follow-up period.



Results ECOG group	Total cost (NOK)	Life years gained	ICER (NOK)
Pembrolizumab progression RWD	1 222 194	2.40	624 458
Pembrolizumab no progression RWD	1 076 043	2.40	456 458

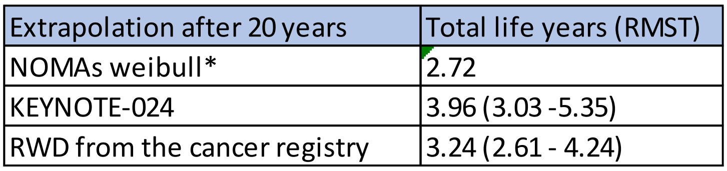
Table 6: ICER of patients with ECOG status ≤ 1



Results ECOG - death group	Total cost (NOK)	Life years gained	ICER (NOK)
Pembrolizumab progression RWD	1 213 611	2.45	578 589
Pembrolizumab no progression RWD	1 064 165	2.45	416 863

Table 7: ICER of patients with ECOG status ≤ 1 and at least 1 month survival

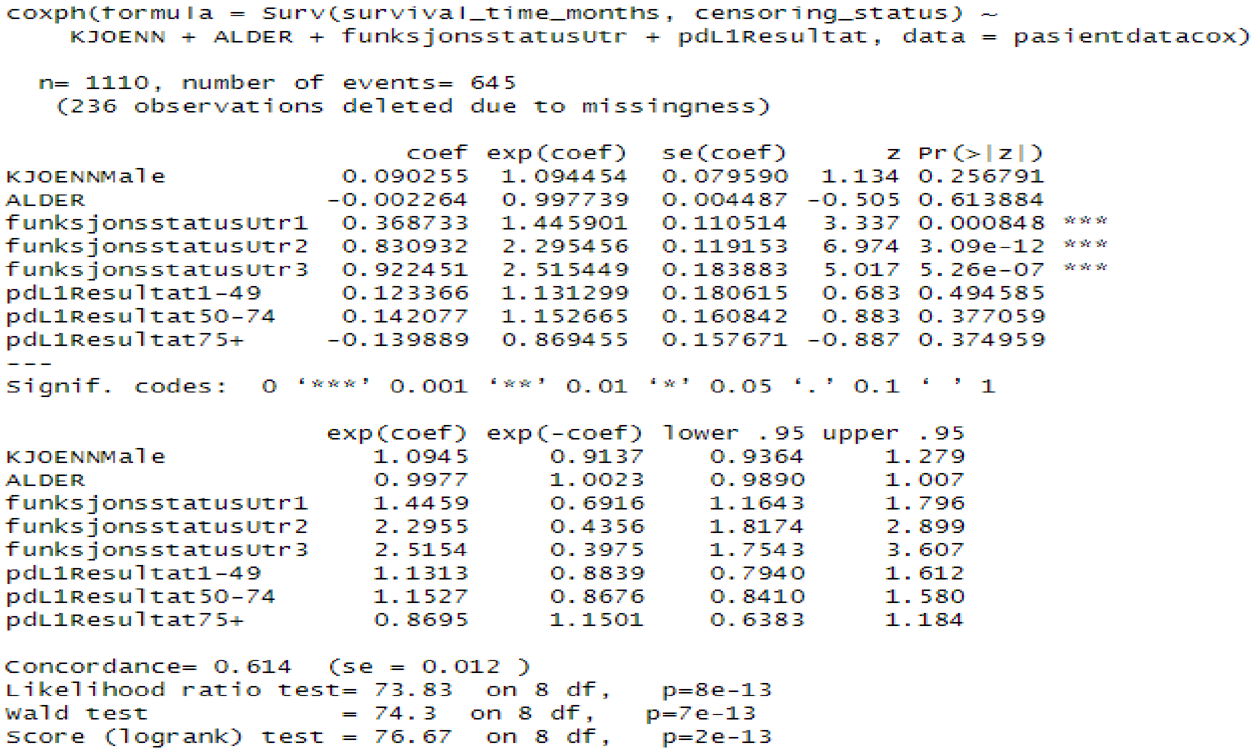
Table 6 and 7 shows how emulating the KEYNOTE-024 may affect the ICER. It is also worth noting that these subgroups are more like the chemotherapy group, but the populations are still not comparable.



Extrapolation after 20 years	Total life years (RMST)
NOMAs weibull*	2.72
KEYNOTE-024	3.96 (3.03 -5.35)
RWD from the cancer registry	3.24 (2.61 - 4.24)

Table 8: Extrapolated RMST of the different datasets. It's important to note that NOMAs extrapolation utilized Weibull, while we utilized a spline model from the survextrap package. The number in parentheses is the range of values, with median as the result.

4.4 Cox regression



```
coxph(formula = Surv(survival_time_months, censoring_status) ~
    KJOENN + ALDER + funksjonsstatusUtr + pdL1Resultat, data = pasientdatacox)

  n= 1110, number of events= 645
   (236 observations deleted due to missingness)

                          coef exp(coef)  se(coef)      z Pr(>|z|)
KJOENNMale            0.090255  1.094454  0.079590  1.134 0.256791
ALDER                -0.002264  0.997739  0.004487 -0.505 0.613884
funksjonsstatusUtr1   0.368733  1.445901  0.110514  3.337 0.000848 ***
funksjonsstatusUtr2   0.830932  2.295456  0.119153  6.974 3.09e-12 ***
funksjonsstatusUtr3   0.922451  2.515449  0.183883  5.017 5.26e-07 ***
pdL1Resultat1-49      0.123366  1.131299  0.180615  0.683 0.494585
pdL1Resultat50-74     0.142077  1.152665  0.160842  0.883 0.377059
pdL1Resultat75+      -0.139889  0.869455  0.157671 -0.887 0.374959
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                    exp(coef) exp(-coef) lower .95 upper .95
KJOENNMale             1.0945     0.9137    0.9364     1.279
ALDER                  0.9977     1.0023    0.9890     1.007
funksjonsstatusUtr1    1.4459     0.6916    1.1643     1.796
funksjonsstatusUtr2    2.2955     0.4356    1.8174     2.899
funksjonsstatusUtr3    2.5154     0.3975    1.7543     3.607
pdL1Resultat1-49       1.1313     0.8839    0.7940     1.612
pdL1Resultat50-74      1.1527     0.8676    0.8410     1.580
pdL1Resultat75+        0.8695     1.1501    0.6383     1.184

Concordance= 0.614  (se = 0.012 )
Likelihood ratio test= 73.83  on 8 df,   p=8e-13
Wald test            = 74.3  on 8 df,   p=7e-13
Score (logrank) test = 76.67  on 8 df,   p=2e-13
```

Figure 4.4.1: Summary of Cox regression model. One patient was removed due to being the only patient with an ECOG level of 4. Baselines are female, ECOG 0 (funksjonstatusUtr1) and PD-L1 1 (PdL1 resultat). Patients with PD-L1 with levels under <1 were combined into a single group to increase statistical power.

While our analysis reveals a statistically significant association between higher ECOG level and mortality, it is crucial to note that the proportional hazards assumption is not fulfilled. This violation, as shown by the p-value of less than 0.05 in the Schoenfeld residuals test, suggests that the impact of ECOG level on survival is inconsistent over time.

4.5 Sensitivity analysis

We preformed multiple sensitivity analyses. Due to some patients being diagnosed several months prior to treatment with pembrolizumab we opted to stratify survival time based on time until initiation of monotherapy after diagnosis, and there was no statistically significant difference.

We also did a sensitivity analysis on patients and date of diagnosis. The patients diagnosed before June 2017 had a significantly shorter survival ($p << 0.01$).

We also compared survival time pre- and post-2019 based on initiation due to the missing data on adjuvant treatment and new treatment guidelines recommending combination therapy as first line treatment. There was no significant difference, but the post-2019 group had a trend of increased survival.

5 Discussion

The HTA conducted by NOMA encountered challenges primarily due to the immature data concerning long-term efficacy. In particular, the extrapolation of survival entailed substantial uncertainty, which was explicitly highlighted by NOMA in their report. Due to the uncertainty combined with the high cost they could not recommend the therapy for the public price (2).

Given our results in chapter 4.2 and 4.3, this section will explore the potential benefits of incorporating real-world data to yield a more precise estimation of long-term survival benefits. Additionally, it will discuss a possible framework for using real world data for temporary approvals.

5.1 Overall survival in the real world and study population

Our median real world OS was 13 (12 - 15). A comparative real world study conducted in the Netherlands (n=83) reported a median OS of 15.8 months (9.4-22.1). One potential explanation for the discrepancy median survival could be attributed to the ECOG status, since 23.6% of our patients had ECOG level 2, and 5% had a ECOG score of 3. In contrast, their population had 3 patients with ECOG level 2 (4%), and no patients with higher ECOG status. Therefore our population is not fully comparable. (39).

Other investigations have also documented higher median OS, although with few or no participants presenting ECOG statuses above 1. We conducted a sensitivity analysis, and by excluding patients with ECOG status above 1 our median OS increased to 17 months (95% CI 15-21), and survival at 36, 48 and 60 months was 34.6%,30.5% and 26.3% respectively. The OS then approach the level found in real world studies on similar patients in USA, which had a median survival of 19.6 months (95% CI 16.6–24.3) (40). Although the median is slightly higher, it is important to note that the other study excluded patients with squamous cell carcinoma, which may explain a higher median OS (38).

Our results are similar to other real-world studies, and especially the subgroup with non-significant difference to KEYNOTE-024 when adjusting for ECOG indicates that the use of real world data may be a valid strategy for evaluating survival.

Table 5,6 and 7 display how changes in selection criteria may have affected the overall survival and cost of treatment. Although the rudimentary adjustment for the prognosis criteria is not fully sound and probably increases overall survival more than the prognosis criteria, it still illustrates the possible effect of adjusting treatment guidelines.

Limiting the real world group to ECOG status ≤ 1 was enough to render the difference between the group and KEYNOTE-024 insignificant, though only barely above the significance level. Longer follow-up is likely to increase the p-value even further and reduce the differences between the groups, based on the tails of the Kaplan-Meier curves.

Survival seems to stabilize after 54 months and the survival curves seem to converge between real world and RCT, but few patients have achieved 54 months of follow-up. We were unable to find any real world data with follow-up beyond 52 months, and unfortunately these studies exclude patients with ECOG status above 1 (41).

Our mean survival is also artificially low due to continuous enrolling of patients.

This is observational data and there are confounding variables that will have an effect on survival. However for pharmacoeconomic assessments also have to consider external validity. Factors such as time on treatment might differ in the real world, which would both affect cost and effectiveness for example (42,43).

We also tried to estimate progression free survival, but had to resort to a proxy, which was switching to a new treatment regimen. Roughly a quarter of patients switched treatment within 24 months, which does not reflect progression free survival in real world data (41).

The subgroup with ECOG status ≥ 2 had a lower median survival that those with ECOG ≤ 1 as shown in fgure 4.2.3 and 4.2.5 (7 vs 17 months). Our results have a higher median survival compared to other real world studies, which report a median survival of 3-4 months (44,45).

These studies however have a lower proportion of patients with ECOG status ≥ 2, and some of the studies did not include patients with ECOG status ≥ 2 (44,45). For example our subgroup had 68 patients (17.5%) with ECOG status of 3, compared to 4 patients (13.8%) (45). Our results also has a larger sample size with longer follow-up. After 24 months our subgroup had roughly three times the overall survival compared to one of the studies, but they only had 2 patients with 24 months of follow-up (44). We could not find any studies on similar subgroups which had more than a handful of patients after 12-24 months, so these results may bring insight on long term overall survival for patients with poor ECOG status.

5.1.1 Extrapolations of overall survival

To extrapolate survival we opted to utilize spline models instead of a parametric Weibull distribution is because they tend to generate the more accurate predictions than other models (46).

We have not evaluated other models, since the extrapolations are not used for further analysis, and it is essential to carefully consider long term extrapolations. When extrapolating immuno-oncology treatments they seem to be more accurate (47,48).

The extrapolated restricted mean OS after 20 years for NOMAs analysis and KEYNOTE-024 was 2.72 vs. 3.96 (3.03 -5.35) years. Real world survival restricted mean OS was 3.96 (3.03 - 5.35) years.

The KEYNOTE-024 spline extrapolation is based on the 5-year update. This illustrates how uncertain extrapolations are based on immature data, and spline models would also likely be inaccurate with immature data (46).

The extrapolations were not used in the pharmacoeconomic assessment but are included to highlight the uncertainty of long-term extrapolations based on immature data. A study which examined extrapolations conducted the French Health Authority, specifically regarding immunotherapy. The study revealed that extrapolations in the HTAs underestimated RMST in 73% of the HTAs, with a mean difference of 13%. One of the extrapolations underestimated the RMST by 43%, or 17 months. It is also worth noting that a ten-year period was extrapolated, instead of 20 years (49).

5.2 Cox regression

The proportional hazard assumption is not met, most likely due to variations in ECOG status over the course of the disease. These changes are not captured in our dataset. Some patients also recorded ECOG status several months before initiating treatment, so it is uncertain if ECOG status unchanged. Changes in ECOG status could significantly affect the prognosis, yet these alterations remain unmeasured and are not accounted for in the analysis. Consequently, while it is valid to assert a statistically significant relationship between higher ECOG status and increased mortality, this association should be interpreted with the understanding that the effect of ECOG status on survival is dynamic and may fluctuate at different stages of the treatment, especially due to disease progression and treatment with chemotherapy (50).

5.3 ICER

Our ICER is quite simplified since we used average dosing and survival. In NOMAs analysis they extrapolated survival, so comparing our results with chemotherapy is not ideal (2). If we extrapolated life years gained would increase, but then we would also have to make assumptions regarding time on treatment. By using averages we yield a conservative results, where we have a lower survival compared to the extrapolated standard of care. Since patients are included continuously, we also have a inflated cost of treatment, since few patients in our receive more than 35 doses, therefore patients with more than 105 weeks of survival have a reduced cost of treatment while contributing to increased survival. This period of reduced cost is often not reached or reduced for our patients due to censoring.

The analysis shows a lower ICER for patients with an ECOG performance status of ≤1, suggesting that treatment restriction based on ECOG may be useful for ensuring the treatment is offered at a cost-effective price point based on local guidelines. Including a requirement of survival prognosis may also further increase the cost-effectiveness of the treatment.

Our findings show a similar ICER within the real would groups even with additional exclusion criteria such as ECOG status ≤ 1, as seen in table 5, 6 and 7. It also shows the possible effect of changing treatment guidelines based on ECOG status.

Imposing restrictions based on ECOG status raises ethical concerns and may inadvertently pressure clinicians to assign lower scores to borderline patients, challenging the balance between cost-effectiveness and equitable care.

The decrease in the ICER is marginal, yet the implications of withholding treatment from approximately one-third of stage IV lung cancer patients are significant. Within our dataset, there is a woman in her early thirties with an ECOG performance level > 1. She was diagnosed early in our follow-up period and is still alive, possibly due to receiving treatment with pembrolizumab.

It raises a critical question: Is it justifiable to exclude her from receiving treatment in favor of patients with an ECOG level ≤ 1, particularly when her potential for a longer life span could surpass that of patients in their 90s with a lower ECOG status? This scenario underscores the ethical and practical challenges of rigidly applying ICER-based restrictions.

Our reasoning for not using extrapolated survival and discounting rates for overall survival is that our simplified and conservative estimations are well within the range of cost-effectiveness, and utilizing extrapolated survival would lower ICER even further. We also aimed to reduce the complexity of our analysis due to the lack of a robust variable describing disease progression.

5.4 Utility and limitations of real world data

Although real-world data serve a purpose when evaluating effectiveness, randomized controlled trials (RCT) remain the gold standard for evaluating efficacy under controlled conditions (51).

However, there are limitations for this study design aswell. The main issue is cost and primary endpoints may develop after decades of follow-up, and a low incidence rate would require a larger study population. This may make it difficult to secure funding, especially for off-patented drugs (52).

The real world is all but controlled and in our thesis we see patients with PD-L1 levels <1% and with ECOG status ≥ 2 receiving monotherapy. Data from KEYNOTE-024 provide no information on treatment effect in these patient groups, and to our knowledge no controlled trials have been conducted on the topic. External validity and information regarding effectiveness for these subgroups would provide crucial insights for physicians, which can be gained by evaluating both real-world data (53).

The benefit-risk may be less favorable outside a clinical trial, when patients deviate from the inclusion criteria in RCTs due to comorbidities or other clinical variables. As shown in our results, real-world effectiveness significantly differ from the clinical trial (figure 4.2.2). Comorbidity in a real-world population may contribute to increased frequency or severity of side effects, further reducing the benefit-risk ratio. Unfortunately, we cannot investigate this due to lack of data on adverse events (54).

5.5 Benefits for HTAs

Our findings suggest that the CRN can effectively be used to estimate survival rates in patients undergoing pembrolizumab monotherapy. By using registry data, we can emulate clinical trials, gauge external validity, and estimate costs based on actual dosing regimens. Although the data may introduce bias in the analysis due to confounding, it may still contribute to pharmacoeconomic assessments. This approach may be a valuable alternative or supplement while waiting for updates from clinical studies, instead of relying on extrapolations from preliminary data.

The use of CRN data could significantly speed up the HTA process by facilitating the use of a revised temporary approval framework, which would still maintain the requirement for plausible cost-effectiveness (55). I will elaborate on my recommendations, which are designed to enhance the efficacy of this approach.

1. Oncology treatments surrounded by uncertainty about their cost-effectiveness should be granted temporary implementation only until definitive evaluations can be conducted.

2. All temporary treatments shall undergo quarterly monitoring through the CRN until sufficient data is amassed to make a conclusive decision regarding their implementation.

3. The variables used for analysis should be standardized, with additional of case-specific variables such as PD-L1 expression.

4. The primary metric for assessing treatment impact should be the years of life gained, with QALYs as a supporting metric.

5. Before the data collection phase, predefined upper/lower thresholds for cost-effectiveness should be set for preliminary analyses.

5.6 Arguments for the temporary approval framework

5.6.1 Point 1 - Uncertainty

All oncology treatments with uncertainty on if it is cost-effective should be temporarily approved. As mentioned, 39% of oncology HTAs have uncertain data, of which 49% are later approved. In general it is only possible to calculate ICER in 58% of HTAs with uncertain data (22).

Considering that 42% of uncertain HTAs have not calculated an ICER, and extrapolations are based on immature data, it is uncertain whether all of these treatments are cost-effective. By ignoring these uncertain implementations, we run the risk of using treatments for years that are cost ineffective. At first, this system should be limited to the treatments with the highest degree of uncertainty regarding cost-effectiveness as a proof of concept, similar to the few temporary approvals already implemented (13,22).

By following up real world data and re-evaluating when data is more certain you can also argue that the initial pharmacoeconomic evaluation should be simplified. Cost estimations with immature data are uncertain due to inaccurate extrapolations (46,56).

One might argue that we should wait until updates from the clinical trial are published, but then we cannot evaluate external validity or treatment effectiveness in subgroups that were excluded, such as patients with ECOG status above 1.

5.6.2 Point 2 - Approval after sufficient data collection

If enough data is collected for an evaluation, a treatment should be approved or denied regardless of a predefined temporary approval period. The rationale for this point hinges on the advantage of formally adopting treatments at once when a comprehensive CEA has been carried out, aiming to diminish the ongoing administrative burden. Swift approval or disapproval of treatments is also advantageous, which may minimize the impact of loss aversion among patients and reducing the duration of uncertainty regarding the potential revocation of access to treatments. This approach can facilitate a more efficient healthcare administration process and may enhance patient well-being by determining treatment availability as soon as possible.

5.6.3 Point 3 - standardization

To streamline the process of data collection, cleaning, and analysis, i suggest that data variables should have the same format across all treatments monitored using CRN data. While recognizing the existence of disease-specific variables, a universal set of variables for all patients, such as ECOG status, smoking history, and level of physical activity—should be incorporated consistently. This standardization provides an opportunity for covariate adjustment, enabling more accurate emulation of target trials (57).

With a standardized dataset, it is possible to apply a uniform analytical template to estimate real-world OS, adjust for common covariates, and perform extrapolation and cost-effectiveness analyses for any treatment. A limitation of utilizing a standardized dataset and values is that the dataset may not capture all relevant details for every case. To avoid this issue additional variables may be necessary to accommodate evaluations with unique considerations.

For instance, transitioning the analysis from pembrolizumab to atezolizumab monotherapy could be done by simply replacing all text with “pembrolizumab” with “atezolizumab” with a search-and-replace function.

Using the same format could enable quarterly analyses, enhancing the number of treatment evaluations possible under temporary approval, which may be useful in certain cases. This would not only shorten the duration for which treatments remain under temporary approval compared to annual reviews and minimize the efforts required for data preparation and code modification.

5.6.4 Point 4 - life years gained

Unfortunately, we could not find literature that examines whether life years gained or QALYs are more appropriate for pharmacoeconomic evaluations when employing real-world data. These opinions are simply based on the issues we faced when trying to estimate QALYs while using a proxy for disease progression.

Life years gained should serve as the primary metric for CEA due to current limitations in the registry. This indicator is less susceptible to confounding variables that may influence QALYs, and it is a hard endpoint. Unlike the complex calculations required for quality-adjusted life QALYs, such as adjustments based on changes in ECOG performance status, life years gained offers a more simplified approach to outcome measurement.

While QALYs remain a valuable supporting metric, it is potentially more prone to variability and inaccuracies due to measurement discrepancies or incomplete data. In our analysis, for instance, using changes in treatment as an indicator of disease progression resulted in only half the percentage of changes compared to those reported in clinical studies. Such discrepancies could significantly affect the estimation of QALYs gained.

A potential solution to these challenges involves the direct incorporation of health-related quality of life metrics into real-world data analysis. However, this approach might be too time-intensive (58).

Quality of life is a part of a ongoing patient population study from 2022, but the response rate is low, which may introduce bias (59).

5.6.5 Point 5 - Threshold for re-evaluation

Implementing automatic regular analyses could benefit from the establishment of threshold values for cost-effectiveness, necessitating data re-evaluation as needed. For instance, should a treatment demonstrate an Incremental Cost-Effectiveness Ratio (ICER) of 1.5 million NOK at the 1-year evaluation point, it would prompt a closer inspection. If the unusual ICER results from data anomalies or errors, the provisional status of the treatment would remain unchanged. Conversely, if initial findings indicate the treatment is indeed cost-effective, a detailed re-assessment would be justified.

Put simply, the predefined upper and lower cost-effectiveness limits act as alerts to re-examine the data. A more thorough examination would be initiated if preliminary results suggest the treatment's cost-effectiveness is significantly lower than initially estimated or surpasses the threshold for cost-effectiveness.

5.6.6 Adverse events

Currently data on adverse events are not available in the CRN registry, but NOMA maintains a registry for side effects. Merging these registries could prove beneficial, potentially providing insight to certain variables that increases the likelihood of adverse events. The advantage of real world data is also the additional benefit of increased participants, which increases the chance for accurately estimating the likelihood of rare adverse events. Additionally, it opens the possibility of uncovering rare adverse events.

The number of adverse effects that remain unreported is unclear. Typically, around five thousand adverse events are reported yearly, of which a thousand are considered severe. In 2021 a total of 523 adverse events were attributed to the ATC group L: antineoplastic and immunomodulating agents (60). Incorporating adverse event reporting into the cancer registry could potentially enhance reporting rates, lead to more precise estimates of their frequency, and possibly aid in developing strategies for prevention.

5.6.7 Automatization and AI

Since variable names and values are usually standardized for the cancer registry, it may be possible to create a semi-finished template for analysis of the real-world data. These templates might be rudimentary and not be fully methodically sound across all diagnoses, but can reduce the workload for statisticians significantly and output a rough first draft that can be utilized to consider a more rigorous analysis.

An even more technologically advanced method may be to utilize AI for these tasks, especially if there already have been several rigorous analyses within the same medical condition. Previous analyses can be utilized as a reference, and methodology can be evaluated afterwards by skilled statisticians and further refined, thus also improving the AI with machine learning (61).

6 Methodological rationale and limitations

6.1 EGFR and ALK mutations

These variables, although most likely measured, were not included in our dataset. Fortunately, we can use H-prescriptions as a proxy, given that the treatments were generally prescribed for these specific mutations.

6.2 Missing regions

Our study population did not include patients treated in Northern Norway (Helse Nord RHF), and some hospitals in the other regions were not included until late 2018. This is may lead to bias, since survival of lung cancer differs between the healthcare regions. According to a recent report, median survival in Norway for all stages of lung cancer was 16.2 months. Patients receiving treatment in northern Norway had a median survival of 14.5 months (62).

There was also a large discrepancy between hospitals within the same healthcare region. Oslo university hospital HF (OUS HF) had a median survival of 22.5 months. Vestre Viken HF was not included in our dataset until 2018, and had a median survival of 16.5 months (62). The cause of the discrepancy may be due to delayed diagnosis and initiation of treatment (63)

Our dataset has no information regarding where patients were treated, so we are unable to investigate if there was an association between overall survival and treatment location.

6.3 Deaths and 2023

We lack data on survival for the year 2023. While we have records of pembrolizumab administration, it’s unclear whether patients passed away or discontinued their treatments after 2022. This gap in data is significant, especially considering 40% of patients were censored. It’s probable that including survival data from 2023 would elevate the average life years gained. This is because pembrolizumab monotherapy ceased being the primary treatment option since October 2020, and there are fewer incident monotherapy patients. Consequently, the patients censored after 2022 are likely high responders with a lower mortality rate. Moreover, patients no longer on treatment may be experiencing complete remission. Such scenarios contribute to an increase in life years gained, achieved at a minimal cost.

6.4 Dates

In our study all dates are set to the 15th of each month for patient anonymity. This adjustment complicates the analysis of treatment initiation. For instance, patients diagnosed on the last day of a month who initiated the treatment the next day are recorded as starting treatment after 30 days. While this tends to balance out over a large group of patients, it nevertheless restricts the precision of our findings and leads to broader confidence intervals in our results.

6.4.1 Cost model

Our cost model is simple compared to NOMAs models since we only used average survival instead of extrapolations. However, the basis for our data is uncertain in respects to disease progression, adverse events, and different patient characteristics between the real world group and chemotherapy study group. Our results are within the limits of cost-effectiveness, we consider the estimate to be satisfactory.

ICER will also likely decrease due to the tail of the Kaplan-Meier survival curve, which will further increase average survival time, but with few or no additional doses of pembrolizumab.

Ideally, we would have a real world chemotherapy as a comparator, and extrapolated both groups with the same model. Comparing matured extrapolated survival with immature extrapolated data is misleading since the immature extrapolation tends to underestimate OS (49).

6.4.2 Combination therapy

We did not construct a separate cost model for patients undergoing combination therapy. We neither attempted to estimate the additional cost of chemotherapy medication, or other costs related to the chemotherapy treatment.

This is due to the results found during our sensitivity analysis regarding time origin. These results indicated that an insignificant number of patients had a history of combination therapy before initiation of monotherapy, and the duration of treatment with combination therapy was short.

The total effect on ICER by calculating additional costs of 1-4 cycles of combination therapy would be insignificant, and highly complicated due to our lack of data regarding adverse events. The cost of pembrolizumab is however included, but not the cost of chemotherapy.

It is also uncertain whether the overall costs of chemotherapy also apply for patients only receiving a couple cycles, such as side effects or supportive treatment. The cost of drug acquisition for chemotherapy is also difficult to calculate due to varying dosages and confidential pricing. This is a minor part of the cost of chemotherapy treatment in total (2).

6.4.3 Switch in treatment regimen

Initially our study design excluded patients who transitioned from pembrolizumab monotherapy to alternative regimens within the first 24 months. However, it became apparent that a significant portion of these patients likely switched therapies due to either disease progression or adverse events. Therefore, we decided to include non-adherent patients in our study group.

This complicates the estimation of cost significantly, since we have no information regarding patient health, degree of progression or adverse events. We have no new information regarding patient characteristics when treatment regimens are modified.

Roughly a quarter of patients switched therapy within 24 months, and approximately half received combination therapy, or chemotherapy alone. Of the patients who switched to combination therapy, some also went on to switch to chemotherapy alone eventually. These switches are in accordance with current guideline recommendations (30).

Due to the diverse treatment regimens and the complex pricing for combination therapies for patients deviating from monotherapy we chose not to calculate the associated costs. This will lower the ICER artificially, but the sensitivity analysis on progression and reducing survival time indicate ICER is within cost-effectiveness.

6.5 Imputation

We did not impute our real world group, since the missing variables frequently overlap, and we were uncertain to the cause of the missingness, since they seemed to be tied to patients with an earlier diagnosis.

6.6 PD-L1 expression

We included patients regardless of PD-L1 expression. The rationale for this choice is that PD-L1 expression was only measured once, and both increases and decreases in expression is common (64). This is especially relevant with patients that were diagnosed with stage I-III cancer, with no information about when they reached stage IV. Pembrolizumab monotherapy is not indicated for patients with PD-L1 expression under 50% as a first line treatment, so we assume treatment regimens were according to guidelines during our follow up period (30).

In patients that were diagnosed with stage IV cancer the proportion of patients that had a PD-L1 expression under 50% was 19.2%. These patients likely started treatment within a couple months despite being registered with a value below the threshold for initiation of monotherapy. It is uncertain whether a new test was conducted, but according to an expert it is unlikely unless there was an overweighing medical need, due to the invasive nature of the required biopsy according to Fagereng, Gro L. (Oncology researcher, meeting 31th January 2024).

The predictive utility of PD-L1 values in our dataset is also dubious, especially considering PD-L1 levels may change in a substantial portion of patients. A study investigating changes in PD-L1 after NSCLC progression reported that approximately 33% of the patients had changes in PD-L1, and 17% had potentially clinically relevant changes. Treatment with chemotherapy significantly increased the likelihood of changes in PD-L1. Changes in expression did not seem to be caused by changes in biopsy location (64).

The patients that had changes in PD-L1 level, 33% encountered changes that pushed their measurements beyond or below the critical threshold of 50%. The study had a follow-up from June 2018 to December 2019, so it might be likely to observe a higher degree of changes in PD-L1 during a longer follow-up period (64).

While PD-L1 serves as a key biomarker to predict response to treatment with immunotherapy, it is not entirely accurate. Other models using additional variables such as ECOG and neutrophil to lymphocyte ratio can be more effective in predicting treatment response (65). Basing treatment decisions solely on PD-L1 levels represents an oversimplification and is not the most effective method in predicting response on an individual level.

6.7 Time origin

We believe that using the first administration of pembrolizumab monotherapy is a valid method. This approach is used in other studies, and for patients who were initially diagnosed at stage III we have no information on when they progressed to stage IV. Consequently, diagnosis date cannot serve as the basis for overall survival (66,67).

It is doubtful whether patients receiving 1-4 cycles of combination therapy will have clinically relevant different survival compared to patients receiving monotherapy (68).

However, by including combination therapy into the survival time we introduce immortal time bias, since the combination therapy patients must survive until they receive monotherapy to be included.

We also conducted a sensitivity analysis, and the effect was insignificant when comparing first dose of monotherapy to first dose of pembrolizumab as time origin. At 3,4 and 5 years, OS increased by less than 0.5%, and median survival did not differ.

we opted not to incorporate the combination therapy into the survival time analysis to mitigate bias and yield more conservative outcomes.

6.8 Lack of real-world data on chemotherapy

Unfortunately, we do not have real world data on chemotherapy alone. This may cause the ICER to increase, since the survival data used by NOMA might benefit from efficacy - effectiveness gap due to selection mechanisms in the clinical study. This may cause the overall survival to be higher in the clinical study than one might expect in a real world setting.

We considered using real world data, but utilizing data before implementation of immunotherapy poses certain risks. Changes in chemotherapy regimens, radiation treatment, and time until diagnosis might differ pre- and post- 2016. There is also risk of cohort effects and period effects.

If we had real world data on chemotherapy from CRN during our follow up period, it is unsure whether the data is useful even when adjusting for cohort and period effects. Since immunotherapy is first line treatment, we would have few remaining patients utilizing chemotherapy, and their characteristics may differ from the pembrolizumab patients.

Using real-world data from countries where chemotherapy remain the primary treatment can be challenging because these countries may rely on chemotherapy due to financial constraints. Real-world data from Norway might not align with data from a developing country, given the potential disparities in healthcare infrastructure, time until diagnosis and socioeconomic factors.

6.9 Disease progression

We have no data on cancer progression. We considered using switch to new treatment regimens as a proxy for progression, but we had significantly lower proportion of patients with disease progression than KEYNOTE-024. At 24 months, we had roughly 25% of patients classified as having disease progression, compared to nearly 50% in KEYNOTE-024 (34).

However clinical disease progression differs from study disease progression, due to increased monitoring and documentation in clinical studies, according to G.L Fagereng, PhD (OUS, 29.01.24).

Due to the major discrepancy, especially also considering the clinical-efficacy gap, we cannot utilize this method as a proxy. It is also not especially useful to utilize in a ICER analysis since the basis of the cost is highly uncertain. we opted to use both the weekly cost of progression and no progression for the whole period to gain a upper and lower limit of cost instead of giving a single value that can be misleading.

6.10 Measurement of variables

All patient characteristics are measured at diagnosis. This complicated the choice of inclusion criteria in our real world group, since patients who might not matching the KEYNOTE-024 inclusion criteria at diagnosis might have been a candidate at a later point.

For the ECOG subgroups nearly a third of the patients may have a higher ECOG status when initiating treatment with pembrolizumab, since they are diagnosed at an earlier cancer stage. They did not initiate treatment until they presumably progressed to stage IV, but we cannot know for certain.

There’s a noticeable overlap between cases with missing ECOG status and those with unspecified cancer, which may be due to the reporting system.

All temporal variables are also rounded to the 15th of every month, which introduces uncertainty.

6.11 Adverse events

There was no information regarding adverse events, this is a limitation in the CRN data.

The most common side effect for pembrolizumab is anemia, which NOMA estimates affects 4.5% of patients. The cost of this adverse event is estimated to be 3114 NOK. NOMA only included costs for adverse events if the adverse event was grade 3 or higher, and with a frequency of at least 5%. Therefore, it is unlikely that lack of data on adverse events for pembrolizumab monotherapy would affect the cost of treatment (2).

6.12 QALYs

Precisely calculating QALYs is hindered by data constraints. This is a setback given the importance of QALYs in pharmacoeconomic evaluations. Although we contemplated employing a proxy for disease progression to approximate QALYs, the inferior quality of this proxy makes it unreliable as a foundation for such estimations. Leveraging inaccurate QALY figures in an ICER which is the basis for pharmacoeconomic assessments, poses considerable risks to financial expenditures and patient well-being.

NOMA also met a variation of this issue in the original assessment, where the reliability of the survival data was questionable due to the short follow-up. Consequently, the drug was deemed not cost-effective under public pricing, largely due to the significant uncertainty surrounding its efficacy and an ICER figure that hovered at the threshold of the pricing limit.

6.13 Current limitations of CRN

Currently, there are limited variables available to adjust for confounding variables or emulate clinical trials.

Details on changes in ECOG status or other factors like cancer stage would increase the precision of pharmacoeconomic evaluations using registry data. However, these variables are only recorded at the time of diagnosis and any additional recording would increase the manual input from physician which could potentially result in lower coverage.

CRN does collect data on additional patient variables as part of a population study. The data may be useful in an pharmacoeconomic assessment; however, the current response rate is low (69).

Conclusion

This investigation serves as a pilot study for integrating real-world data into pharmacoeconomic assessments of treatments with immature data within the New methods system.

Currently real world data from CRN can be used for pharmacoeconomic assessments based on life years gained. Our estimations indicate an ICER of 474 621 - 689 641 NOK, compared to 784 851 NOK in NOMAs assessment.

Unfortunately, the real world data currently lacks robust variables for calculating an ICER, and further development and additional variables would be beneficial to improve accuracy of cost-effectiveness estimates, and estimating cost-effectiveness based on QALYs.

We have suggested a framework that can be used for validating long term extrapolations based on immature data, and we have identified that ECOG status is the main variable associated with the efficacy-effectiveness gap in overall survival.

References

1. Helse- og omsorgsdepartementet. Meld. St. 34 (2015–2016) - Verdier i pasientens helsetjeneste — Melding om prioritering [Internet]. Regjeringen.no. https://www.regjeringen.no/no/dokumenter/meld.-st.-34-20152016/id2502758/; 2016 [cited 2023 Nov 7]. Available from: https://www.regjeringen.no/no/dokumenter/meld.-st.-34-20152016/id2502758/

2. Strøm B oddvar, Sagdahl E, Endersen H. Pembrolizumab (Keytruda) til førstelinjebehandling av metastatisk ikke-småcellet lungekreft med tumor som uttrykker PD-L1 med ≥ 50 %. [Internet]. Statens legemiddelverk; 2017 Sep [cited 2023 Aug 28]. Report No.: ID2016_067. Available from: https://nyemetoder.no/Documents/Rapporter/Pembrolizumab%20-HTA.pdf

3. Nye Metoder. Bakgrunn: Hvorfor har vi Nye metoder? [Internet]. https://nyemetoder.no/om-systemet/bakgrunn-hvorfor-har-vi-nye-metoder; [cited 2023 Nov 7]. Available from: https://nyemetoder.no/om-systemet/bakgrunn-hvorfor-har-vi-nye-metoder

4. Nye metoder. Forslag om nasjonale metodevurderinger [Internet]. https://nyemetoder.no/om-systemet/forslag-om-nasjonale-metodevurderinger; [cited 2023 Nov 8]. Available from: https://nyemetoder.no/om-systemet/forslag-om-nasjonale-metodevurderinger

5. Nye metoder. Bestillerforum for nye metoder [Internet]. https://www.nyemetoder.no/om-systemet/bestillerforum-for-nye-metoder/; [cited 2024 Mar 7]. Available from: https://www.nyemetoder.no/om-systemet/bestillerforum-for-nye-metoder/

6. Nye metoder. Metodevurderinger [Internet]. https://nyemetoder.no/om-systemet/metodevurderinger; [cited 2023 Nov 7]. Available from: https://nyemetoder.no/om-systemet/metodevurderinger

7. Nye metoder. Faser ved innføring av nye metoder - implementering [Internet]. https://nyemetoder.no/implementering; [cited 2023 Nov 12]. Available from: https://nyemetoder.no/implementering

8. Nye metoder. Innføring av nye legemidler: Slik var tidsbruken i 2022 [Internet]. https://www.nyemetoder.no/aktuelt/innforing-av-nye-legemidler-slik-var-tidsbruken-i-2022; 13.03.23 [cited 2024 Mar 7]. Available from: https://www.nyemetoder.no/aktuelt/innforing-av-nye-legemidler-slik-var-tidsbruken-i-2022

9. Sykehusinnkjøp. Første legemidler godkjent i nytt hurtigløp for immunterapi [Internet]. https://www.sykehusinnkjop.no/nyheter/nyheter-2024/nye-metoder-hurtiglop-kreft/; 12.03.24 [cited 2024 Mar 7]. Available from: https://www.sykehusinnkjop.no/nyheter/nyheter-2024/nye-metoder-hurtiglop-kreft/

10. Strøm B oddvar, Krontveit R, Michel YA. Entrektinib (Rozlytrek) - Indikasjon II [Internet]. Statens legemiddelverk; 2021 Feb [cited 2024 Mar 4]. Report No.: ID2019_119. Available from: https://www.nyemetoder.no/metoder/entrektinib-rozlytrek-indikasjon-ii/

11. Urbaniak A, Vidal C, Røshol H. Gilteritinib (Xospata) [Internet]. Statens legemiddelverk; 2020 Feb [cited 2024 Mar 5]. Report No.: ID2019_095. Available from: https://www.nyemetoder.no/metoder/gilteritinib-xospata/

12. Grøvan A, Michel YA, Eriksen HM, Hjelme K. Voretigene Neparvovec (Luxturna) - Genterapi ved Lebers medfødte synstap knyttet til mutasjoner i RPE65 genet [Internet]. Statens legemiddelverk; 2020 Jul [cited 2024 Mar 5]. Report No.: ID2016_057. Available from: https://www.nyemetoder.no/metoder/voretigene-neparvovec-luxturna/

13. Kalveland J. Disse 14 medisinene har sykehusene tatt i bruk gjennom <<alternative>> avtaler [Internet]. https://www.dagensmedisin.no/beslutningsforum-cystisk-fibrose-hjerteamyloidos/disse-14-medisinene-har-sykehusene-tatt-i-bruk-gjennom-alternative-avtaler/577479; 2023 [cited 2024 Mar 5]. Available from: https://www.dagensmedisin.no/beslutningsforum-cystisk-fibrose-hjerteamyloidos/disse-14-medisinene-har-sykehusene-tatt-i-bruk-gjennom-alternative-avtaler/577479

14. Beslutningsforum. Rammeverk for prisavtaler Gjeldende fra 23. Juni 2020 [Internet]. Nye metoder; 2020. Available from: https://www.sykehusinnkjop.no/49614b/siteassets/nyheter/beslutningsfourm-22.-juni-2020/rammeverk-prisavtaler-besluttet-22juni2020.pdf

15. Wenzl M, Chapman S. Performance-based managed entry agreements for new medicines in OECD countries and EU member states: How they work and possible

improvements going forward [Internet]. Paris: OECD; 2019 Dec [cited 2024 May 9]. Available from: https://www.oecd-ilibrary.org/social-issues-migration-health/performance-based-managed-entry-agreements-for-new-medicines-in-oecd-countries-and-eu-member-states_6e5e4c0f-en

16. Kreftregisteret. Lungekreft [Internet]. Lungekreft. https://www.kreftregisteret.no/Temasider/kreftformer/Lungekreft/; 06.09.23 [cited 2023 Aug 28]. Available from: https://www.kreftregisteret.no/Temasider/kreftformer/Lungekreft/

17. Reck M, Rodríguez-Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A, et al. Pembrolizumab versus Chemotherapy for PD-L1–Positive Non–Small-Cell Lung Cancer. New England Journal of Medicine [Internet]. 2016 Nov [cited 2023 Sep 7];375(19):1823–33. Available from: https://doi.org/10.1056/NEJMoa1606774

18. ACRIN Cancer Research Group. ECOG Performance Status Scale [Internet]. https://ecog-acrin.org/resources/ecog-performance-status/; [cited 2024 May 9]. Available from: https://ecog-acrin.org/resources/ecog-performance-status/

19. UpToDate. Eastern Cooperative Oncology Group (ECOG) performance status [Internet]. https://www.uptodate.com/contents/image?imageKey=HEME/72901; [cited 2024 May 9]. Available from: https://www.uptodate.com/contents/image?imageKey=HEME/72901

20. Tai TA, Latimer NR, Benedict Á, Kiss Z, Nikolaou A. Prevalence of Immature Survival Data for Anti-Cancer Drugs Presented to the National Institute for Health and Care Excellence and Impact on Decision Making. Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research. 2021 Apr;24(4):505–12.

21. Kang J, Cairns J. "Don't Think Twice, It's All Right": Using Additional Data to Reduce Uncertainty Regarding Oncologic Drugs Provided Through Managed Access Agreements in England. PharmacoEconomics - Open [Internet]. 2023 Jan [cited 2024 Feb 28];7(1):77–91. Available from: https://doi.org/10.1007/s41669-022-00369-9

22. Fagereng GL, Morvik AM, Reinvik Ulimoen S, Ringerud AM, Dahlen Syversen I, Sagdahl E. The impact of level of documentation on the accessibility and affordability of new drugs in Norway. Frontiers in Pharmacology [Internet]. 2024 [cited 2024 Mar 1];15. Available from: https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2024.1338541

23. National Institute for Health and Care Excellence. Technology appraisal data: Cancer appraisal recommendations [Internet]. https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-technology-appraisal-guidance/data/cancer-appraisal-recommendations; [cited 2024 Feb 29]. Available from: https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-technology-appraisal-guidance/data/cancer-appraisal-recommendations

24. Enerly E, Holmstrøm L, Knudsen KO, Skog A, Heien IH. INSPIRE:lungekreft. Evaluering av pilotprosjekt [Internet]. Oslo: Kreftregisteret; 2021 [cited 2023 Dec 13]. Available from: https://www.kreftregisteret.no/globalassets/publikasjoner-og-rapporter/inspire/inspire_lungekreft_evaluering-av-pilotprosjekt.pdf

25. Zappa C, Mousa SA. Non-small cell lung cancer: Current treatment and future advances. Translational Lung Cancer Research [Internet]. 2016 Jun [cited 2023 Oct 29];5(3):288–300. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4931124/

26. Kwok G, Yau TCC, Chiu JW, Tse E, Kwong YL. Pembrolizumab (Keytruda). Human Vaccines & Immunotherapeutics. 2016 Nov;12(11):2777–89.

27. Norum J, Antonsen MA, Tollåli T, Al-Shibli K, Andersen G, Svanqvist KH, et al. Pembrolizumab as second-line therapy in non-small cell lung cancer in northern Norway: Budget impact and expected gain-a model-based analysis. ESMO open. 2017;2(3):e000222.

28. Sousa GF de, Wlodarczyk SR, Monteiro G. Carboplatin: Molecular mechanisms of action associated with chemoresistance. Brazilian Journal of Pharmaceutical Sciences [Internet]. 2014-Oct-Dec [cited 2023 Oct 30];50:693–701. Available from: https://www.scielo.br/j/bjps/a/9F6tZpxsm7spMKjGn6Z3kvr/

29. Capasso A. Vinorelbine in cancer therapy. Current Drug Targets. 2012 Jul;13(8):1065–71.

30. Helsedirektoratet. Lungekreft, mesoteliom og thymom - handlingsprogram [Internet]. https://www.helsedirektoratet.no/retningslinjer/lungekreft-mesoteliom-og-thymom-handlingsprogram/ikke-kurativ-behandling-av-ikke-smacellet-lungekreft; [cited 2023 Oct 14]. Available from: https://www.helsedirektoratet.no/retningslinjer/lungekreft-mesoteliom-og-thymom-handlingsprogram/ikke-kurativ-behandling-av-ikke-smacellet-lungekreft

31. Giuliani J. Cost-Effectiveness of Pembrolizumab in Combination with Platinum-Based Chemotherapy in First-Line for Squamous and Nonsquamous Non Small-Cell Lung Cancer. Far from Economic Sustainability. Eurasian Journal of Medicine and Oncology [Internet]. 2021 [cited 2023 Aug 27]; Available from: https://ejmo.org/10.14744/ejmo.2021.96518/

32. Zhang C, Zhang J, Tan J, Tian P, Li W. Cost-Effectiveness of Pembrolizumab for the treatment of Non–Small-Cell lung cancer: A systematic review. Frontiers in Oncology [Internet]. 2022 Aug [cited 2023 Aug 28];12:815587. Available from: https://www.frontiersin.org/articles/10.3389/fonc.2022.815587/full

33. Jackson CH. Survextrap: A package for flexible and transparent survival extrapolation. BMC Medical Research Methodology [Internet]. 2023 Nov [cited 2024 Feb 21];23(1):282. Available from: https://doi.org/10.1186/s12874-023-02094-1

34. Reck M, Rodríguez-Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A, et al. Five-Year Outcomes With Pembrolizumab Versus Chemotherapy for Metastatic Non-Small-Cell Lung Cancer With PD-L1 Tumor Proportion Score ≥ 50. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology. 2021 Jul;39(21):2339–49.

35. Magirr D. Enhanced Kaplan-Meier Curves [Internet]. https://dominicmagirr.shinyapps.io/enhanceKM/; [cited 2024 Jan 10]. Available from: https://dominicmagirr.shinyapps.io/enhanceKM/

36. Rohatgi A. WebPlotDigitizer 4.6 [Internet]. https://automeris.io/WebPlotDigitizer/; [cited 2024 Jan 10]. Available from: https://automeris.io/WebPlotDigitizer/

37. Hao B, Li F, Wan X, Pan S, Li D, Song C, et al. Squamous cell carcinoma predicts worse prognosis than adenocarcinoma in stage IA lung cancer patients: A population-based propensity score matching analysis. Frontiers in Surgery [Internet]. 2022 Aug [cited 2024 Apr 16];9:944032. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9461700/

38. Wang BY, Huang JY, Chen HC, Lin CH, Lin SH, Hung WH, et al. The comparison between adenocarcinoma and squamous cell carcinoma in lung cancer patients. Journal of Cancer Research and Clinical Oncology. 2020 Jan;146(1):43–52.

39. Cramer-van der Welle CM, Verschueren MV, Tonn M, Peters BJM, Schramel FMNH, Klungel OH, et al. Real-world outcomes versus clinical trial results of immunotherapy in

stage IV non-small cell lung cancer (NSCLC) in the Netherlands. Scientific Reports [Internet]. 2021 Mar [cited 2024 Jan 24];11(1):6306. Available from: https://www.nature.com/articles/s41598-021-85696-3

40. Velcheti V, Hu X, Yang L, Pietanza MC, Burke T. Long-Term Real-World Outcomes of First-Line Pembrolizumab Monotherapy for Metastatic Non-Small Cell Lung Cancer With ≥50% Expression of Programmed Cell Death-Ligand 1. Frontiers in Oncology [Internet]. 2022 [cited 2024 Jan 29];12. Available from: https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2022.834761

41. Velcheti V, Hu X, Yang L, Pietanza MC, Burke T. Long-Term Real-World Outcomes of First-Line Pembrolizumab Monotherapy for Metastatic Non-Small Cell Lung Cancer With ≥50% Expression of Programmed Cell Death-Ligand 1. Frontiers in Oncology [Internet]. 2022 [cited 2024 Jan 24];12. Available from: https://www.frontiersin.org/articles/10.3389/fonc.2022.834761

42. Roberts MH, Ferguson GT. Real-World Evidence: Bridging Gaps in Evidence to Guide Payer Decisions. PharmacoEconomics - Open [Internet]. 2021 Mar [cited 2024 Mar 29];5(1):3–11. Available from: https://doi.org/10.1007/s41669-020-00221-y

43. Velcheti V, Hu X, Li Y, El-Osta H, Pietanza MC, Burke T. Real-World Time on Treatment with First-Line Pembrolizumab Monotherapy for Advanced NSCLC with PD-L1 Expression ≥ 50%: 3-Year Follow-Up Data. Cancers [Internet]. 2022 Feb [cited 2024 Mar 29];14(4):1041. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8870405/

44. Facchinetti F, Mazzaschi G, Barbieri F, Passiglia F, Mazzoni F, Berardi R, et al. First-line pembrolizumab in advanced non-small cell lung cancer patients with poor performance status. European Journal of Cancer (Oxford, England: 1990). 2020 May;130:155–67.

45. Sehgal K, Gill RR, Widick P, Bindal P, McDonald DC, Shea M, et al. Association of Performance Status With Survival in Patients With Advanced Non–Small Cell Lung Cancer Treated With Pembrolizumab Monotherapy. JAMA Network Open [Internet]. 2021 Feb [cited 2024 May 9];4(2):e2037120. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7879233/

46. Gray J, Sullivan T, Latimer NR, Salter A, Sorich MJ, Ward RL, et al. Extrapolation of Survival Curves Using Standard Parametric Models and Flexible Parametric Spline Models:

Comparisons in Large Registry Cohorts with Advanced Cancer. Medical Decision Making [Internet]. 2021 Feb [cited 2024 Feb 16];41(2):179–93. Available from: https://doi.org/10.1177/0272989X20978958

47. Bullement A, Willis A, Amin A, Schlichting M, Hatswell AJ, Bharmal M. Evaluation of survival extrapolation in immuno-oncology using multiple pre-planned data cuts: Learnings to aid in model selection. BMC Medical Research Methodology [Internet]. 2020 May [cited 2024 Feb 16];20(1):103. Available from: https://doi.org/10.1186/s12874-020-00997-x

48. Shao T, Zhao M, Liang L, Shi L, Tang W. Impact of Extrapolation Model Choices on the Structural Uncertainty in Economic Evaluations for Cancer Immunotherapy: A Case Study of Checkmate 067. PharmacoEconomics - Open [Internet]. 2023 May [cited 2024 May 7];7(3):383–92. Available from: https://doi.org/10.1007/s41669-023-00391-5

49. Grumberg V, Roze S, Chevalier J, Borrill J, Gaudin AF, Branchoux S. A Review of Overall Survival Extrapolations of Immune-Checkpoint Inhibitors Used in Health Technology Assessments by the French Health Authorities. International Journal of Technology Assessment in Health Care. 2022 Mar;38(1):e28.

50. Käsmann L, Taugner J, Eze C, Roengvoraphoj O, Dantes M, Gennen K, et al. Performance Status and Its Changes Predict Outcome for Patients With Inoperable Stage III NSCLC Undergoing Multimodal Treatment. Anticancer Research [Internet]. 2019 Sep [cited 2024 Mar 29];39(9):5077–81. Available from: https://ar.iiarjournals.org/content/39/9/5077

51. Rationale, Strengths, and Limitations of Real-World Evidence in Oncology: A Canadian Review and Perspective The Oncologist Oxford Academic [Internet]. https://academic.oup.com/oncolo/article/27/9/e731/6619081; [cited 2024 Feb 15]. Available from: https://academic.oup.com/oncolo/article/27/9/e731/6619081

52. Sanson-Fisher RW, Bonevski B, Green LW, D’Este C. Limitations of the Randomized Controlled Trial in Evaluating Population-Based Health Interventions. American Journal of Preventive Medicine [Internet]. 2007 Aug [cited 2024 Feb 15];33(2):155–61. Available from: https://www.sciencedirect.com/science/article/pii/S0749379707002255

53. Persaud N, Mamdani MM. External validity: The neglected dimension in evidence ranking. Journal of Evaluation in Clinical Practice. 2006 Aug;12(4):450–3.

54. Eichler HG, Abadie E, Breckenridge A, Flamion B, Gustafsson LL, Leufkens H, et al. Bridging the efficacy–effectiveness gap: A regulator's perspective on addressing variability of drug response. Nature Reviews Drug Discovery [Internet]. 2011 Jul [cited 2024 Feb 15];10(7):495–506. Available from: https://www.nature.com/articles/nrd3501

55. Kortversjon sluttraport - Utredning og implementering av ordninger for midlertidig innføring og revurdering av nye metoder i nye metoder for å tilrettelegge for innføring av persontilpasset medisin i tjenesten [Internet]. Helse Vest; 2021 Jun. Available from: https://www.helse-vest.no/499f95/contentassets/736c150c5692409e813e68c414797309/styredokument-2021/01.09.2021/sak-06921-vedlegg-1---kortversjon-sluttrapport---oppdrag-midlertidig-innforing-og-revurdering-i-nye-metoder-24.pdf.pdf

56. van Oostrum I, Ouwens M, Remiro-Azócar A, Baio G, Postma MJ, Buskens E, et al. Comparison of Parametric Survival Extrapolation Approaches Incorporating General Population Mortality for Adequate Health Technology Assessment of New Oncology Drugs. Value in Health [Internet]. 2021 Sep [cited 2024 Mar 29];24(9):1294–301. Available from: https://www.sciencedirect.com/science/article/pii/S1098301521001698

57. Kwee SA, Wong LL, Ludema C, Deng CK, Taira D, Seto T, et al. Target Trial Emulation: A Design Tool for Cancer Clinical Trials. JCO Clinical Cancer Informatics [Internet]. 2023 Sep [cited 2024 Mar 17];(7):e2200140. Available from: https://ascopubs.org/doi/full/10.1200/CCI.22.00140

58. Quality of Life Measure - an overview ScienceDirect Topics [Internet]. https://www.sciencedirect.com/topics/medicine-and-dentistry/quality-of-life-measure; [cited 2024 Mar 17]. Available from: https://www.sciencedirect.com/topics/medicine-and-dentistry/quality-of-life-measure

59. Fjellbirkeland L, Wahl SGF, Haram PM, Helbekkmo N, Helland Å, Majak PP, et al. Årsrapport for lungekreft 2023 [Internet]. Kreftregisteret; 07.05.24 [cited 2024 May 8]. Available from: https://www.kreftregisteret.no/Generelt/Rapporter/Arsrapport-fra-kvalitetsregistrene/Arsrapport-for-lungekreft/arsrapport-for-lungekreft-2023/

60. NOMA. Årsrapport om bivirkninger 2021 [Internet]. https://www.dmp.no/bivirkninger-og-sikkerhet/bivirkningsrapporter-og-

oversikter/arsrapporter-bivirkninger/arsrapport-om-bivirkninger-2021; 2024 [cited 2024 May 8]. Available from: https://www.dmp.no/bivirkninger-og-sikkerhet/bivirkningsrapporter-og-oversikter/arsrapporter-bivirkninger/arsrapport-om-bivirkninger-2021

61. Liu F, Panagiotakos D. Real-world data: A brief review of the methods, applications, challenges and opportunities. BMC Medical Research Methodology [Internet]. 2022 Nov [cited 2024 Feb 1];22(1):287. Available from: https://doi.org/10.1186/s12874-022-01768-6

62. Våg I, Fjellbirkeland L, Solberg SK, Helland Å, Wahl SGF, Haram PM, et al. Årsrapport 2022 med resultater og forbedringstiltak fra Nasjonalt kvalitetsregister for lungekreft. [Internet]. Oslo: Kreftregisteret; 10.05.23. Available from: https://www.kreftregisteret.no/Generelt/Rapporter/Arsrapport-fra-kvalitetsregistrene/Arsrapport-for-lungekreft/arsrapport-for-lungekreft-2022/

63. Johansen E. Helseministeren mener kreftpasienter i Nord-Norge får dårligere behandling [Internet]. NRK. https://www.nrk.no/tromsogfinnmark/helseministeren-mener-kreftpasienter-i-nord-norge-far-darligere-behandling-1.16629363; 2023 [cited 2024 May 9]. Available from: https://www.nrk.no/tromsogfinnmark/helseministeren-mener-kreftpasienter-i-nord-norge-far-darligere-behandling-1.16629363

64. Frank MS, Bødtger U, Høegholm A, Stamp IM, Gehl J. Re-biopsy after first line treatment in advanced NSCLC can reveal changes in PD-L1 expression. Lung Cancer [Internet]. 2020 Nov [cited 2024 Jan 24];149:23–32. Available from: https://www.sciencedirect.com/science/article/pii/S0169500220305948

65. Prelaj A, Boeri M, Robuschi A, Ferrara R, Proto C, Lo Russo G, et al. Machine Learning Using Real-World and Translational Data to Improve Treatment Selection for NSCLC Patients Treated with Immunotherapy. Cancers [Internet]. 2022 Jan [cited 2024 Feb 1];14(2):435. Available from: https://www.mdpi.com/2072-6694/14/2/435

66. Welle CMC der, Peters BJM, Schramel FMNH, Klungel OH, Groen HJM, Garde EMW van de. Systematic evaluation of the efficacy–effectiveness gap of systemic treatments in metastatic nonsmall cell lung cancer. European Respiratory Journal [Internet]. 2018 Dec [cited 2024 Feb 7];52(6). Available from: https://erj.ersjournals.com/content/52/6/1801100

67. Terai H, Soejima K, Shimokawa A, Horinouchi H, Shimizu J, Hase T, et al. Real-World Data Analysis of Pembrolizumab Monotherapy for NSCLC Using Japanese

Postmarketing All-Case Surveillance Data. JTO clinical and research reports. 2022 Nov;3(11):100404.

68. Matsumoto H, Kobayashi N, Somekawa K, Fukuda N, Kaneko A, Kamimaki C, et al. Pembrolizumab monotherapy versus pembrolizumab plus chemotherapy in patients with non-small-cell lung cancer: A multicenter retrospective trial. Thoracic Cancer. 2022 Jan;13(2):228–35.

69. Kreftregisteret. Kvalitetsmål for Lungekreftregisteret [Internet]. https://www.kreftregisteret.no/Registrene/Kvalitetsregistrene/Kvalitetsregister-for-lungekreft/Kvalitetsmal/; 2023 [cited 2024 Apr 14]. Available from: https://www.kreftregisteret.no/Registrene/Kvalitetsregistrene/Kvalitetsregister-for-lungekreft/Kvalitetsmal/

Appendix

This first appendix includes the chunks of R-code that were used while working on this project, including my notes.

D:/masteroppgaven backup/raw data/0 inputtering/Med avvik.Rmd C:/masteroppgaven lokal/raw data/0 inputtering/

r ref.label=‘blabla2’, results=‘hide’, echo = TRUE}

--- title: “rens data 1” author: “jeroen” date: “2023-11-15” output: html_document ---

```
knitr::opts_chunk$set(echo = TRUE)
# Detect the computer's name
computer_name <- Sys.info()["nodename"]

# Set the working directory based on the computer's name
if (computer_name == "JEROEN-LAPTOP") {
  setwd("C:\\masteroppgaven lokal\\raw data\\0 inputtering")
} else if (computer_name == "JEROENHAUKAAS") {
  setwd("D:/Masteroppgaven backup/raw data/0 inputtering")
} else {
  stop("Unknown computer: unable to set the working directory")
}

# Load data
pasientdata <- read.csv( "realworldadherentmedavvik.csv", header = TRUE, sep = ",")
sykehusdata <- read.csv("sykehusadherentmedavvik.csv", header = TRUE, sep = ",")
straaledata <- read.csv("Utlevert_straaledata_4082.csv", header = TRUE, sep = ";")
exclusion_log <- read.csv("exclusion_log2.csv")


library(tidyverse)
library(ggplot2)
library(janitor)
library(mice)
library(dplyr)
```

```r
library(DataExplorer)
library(webshot2)
library(tidyverse)
library(tidyr)
library(janitor)
library(rstatix)
library(remotes)
library(kableExtra)
library(devtools)
library(report)
library(sjPlot)
#library(ggstatsplot)
library(survival)
library(survminer)
library(biostat3)
library(tidyverse)
library(ggsurvfit)
library(dplyr)
library(gtsummary)
library(gridExtra)
library(scales)
```

#Oversetter variabler ## Legger inn sykehusdata

```r
sykehusdata_unique <- sykehusdata %>%
                      arrange(first_dose) %>%
                      group_by(PID) %>%
                      slice(1)




# Merge the datasets
pasientdata <- merge(pasientdata, sykehusdata_unique[, c("PID", "first_dose")], by = "PID",
all.x = FALSE)
```

7.1 legger inn survivaltime

```
censoring_date <- as.Date("2023-01-01")
pasientdata <- pasientdata %>%
  mutate(
    DIAGNOSEDATO = as.Date(DIAGNOSEDATO, format = "%d%b%Y"),
    STATUSDATO = as.Date(STATUSDATO, format = "%d%b%Y"),
    first_dose = as.Date(first_dose),
    # Calculate survival time, using censoring_date for those still alive
    SurvivalTime = if_else(is.na(STATUSDATO),
                          as.numeric(censoring_date - DIAGNOSEDATO),
                          as.numeric(STATUSDATO - DIAGNOSEDATO)))
```

7.2 legger inn tid til første dose

```
pasientdata <- pasientdata %>%
            mutate(time_until_first_dose = first_dose - DIAGNOSEDATO)
```

7.3 undersøker de som har flere sykdommstilfeller

```
patients_with_multiple_SIDs <- pasientdata %>%
                              group_by(PID) %>%
                              summarise(distinct_SIDs = n_distinct(SID)) %>%
                              filter(distinct_SIDs > 1)
```

#Fjerning av pasienter mange urelaterte sykdommer, fjerner SID som er lengst unna SID. Fjerner også pasienter som begynner veldig sent pga at legemiddelet ikke har blitt innført enda.

##stratifisering pre, post beslutningsforum vedtak

```
censoring_date <- as.Date("2022-12-31")

# Calculate Survival Time
pasientdata$survival_time <- ifelse(is.na(pasientdata$STATUSDATO),
                             as.numeric(difftime(censoring_date, pasientdata$first_dose, units = "days")),
                             as.numeric(difftime(pasientdata$STATUSDATO, pasientdata$first_d
```

```r
ose, units = "days")))

# Convert Survival Time to months and round to nearest month
pasientdata$survival_time_months <- round(pasientdata$survival_time / 30.44)

# Censoring status: 0 if alive (censored), 1 if dead (event occurred)
pasientdata$censoring_status <- ifelse(is.na(pasientdata$STATUSDATO), 0, 1)
# Create a new column for stratification based on June 2017
pasientdata$pre_post_June_2017 <- ifelse(pasientdata$DIAGNOSEDATO < as.Date("2017-06-01"), "Pre-June 2017", "Post-June 2017")

# Create the survival object using months with stratification
surv_obj_stratified <- Surv(time = pasientdata$survival_time_months, event = pasientdata$censoring_status)

# Fit the Kaplan-Meier survival curve with stratification
km_fit_stratified <- survfit(surv_obj_stratified ~ pre_post_June_2017, data = pasientdata)

# Generate the survival plot with stratification
survival_plot_stratified <- ggsurvplot(
  km_fit_stratified, data = pasientdata, conf.int = TRUE,
  risk.table = TRUE,
  xlab = "Months since first dose", ylab = "Survival probability",
  title = "Kaplan-Meier Survival Curve by Pre and Post June 2017",
  ggtheme = theme_minimal(),
  break.x.by = 3  # Set x-axis breaks every 3 months
)
median_survival_times <- surv_median(km_fit_stratified)
print(median_survival_times)

# Perform the log-rank test
log_rank_test <- survdiff(surv_obj_stratified ~ pre_post_June_2017, data = pasientdata)

# Extract the p-value from the log-rank test result (keep in scientific notation)
```

```
log_rank_p_value <- 1 - pchisq(log_rank_test$chisq, length(log_rank_test$n) - 1)

# Print the results of the log-rank test
print(log_rank_test)

# Extract median survival times
median_survival_times <- surv_median(km_fit_stratified)
print(median_survival_times)

# Check if median survival times were extracted correctly
if (length(median_survival_times$surv.median) == 0) {
  # Create a placeholder for median survival times if not extracted
  median_survival_times$surv.median <- rep(NA, length(median_survival_times$strata))
}

# Combine median survival times and log-rank p-value into one data frame
combined_results <- data.frame(
  Strata = median_survival_times$strata,
  MedianSurvival = median_survival_times$surv.median,
  LogRankPValue = rep(format(log_rank_p_value, scientific = TRUE), length(median_survival_times$strata))
)

# Print and save the combined results
print(combined_results)

# Save the combined results as a CSV file (or any other format you prefer)
write.csv(combined_results, "CombinedSurvivalAndLogRankResults.csv", row.names = FALSE)

# Print and save the stratified plot
print(survival_plot_stratified)
```

```
ggsave("KMStratifiedPrePostJune2017.png", survival_plot_stratified$plot, width = 15, height = 11, bg = "#FFFDFB", dpi = 300)

before_exclusion_count <- length(unique(pasientdata$PID))
# List of SIDs to remove
sids_to_remove <- c(958, 875, 2295, 1066, 1739, 2420, 1535, 1493, 3374, 2681, 3081, 526, 296,1387,3446,3237,1563,3125,3136)

# Remove records with the specified SIDs
pasientdata <- pasientdata %>%
                    filter(!SID %in% sids_to_remove)



after_exclusion_count <- length(unique(pasientdata$PID))

# Update log
exclusion_log <- rbind(exclusion_log, data.frame(
  Step = 6,
  Reason = "No outliers",
  ExcludedCount = before_exclusion_count - after_exclusion_count,
  RemainingCount = after_exclusion_count
))
```

##fjerner alle som fikk diagnosen før juni 2017 (super signifikant forskjell med pverdi på 0,006)

Levetid Median Øvre/nedre KI pre_post_June_2017=Post-June 2017 12 10 15
pre_post_June_2017=Pre-June 2017 5 3 8

```
before_exclusion_count <- length(unique(pasientdata$PID))
pasientdata <- pasientdata %>%
  filter(DIAGNOSEDATO >= as.Date("2017-06-01") & DIAGNOSEDATO <= as.Date("2022-12-31"))
after_exclusion_count <- length(unique(pasientdata$PID))
sids_to_remove <- c(43,456,778,993,2174,2175,2218,2851,3261)
```

```
# Remove records with the specified SIDs
pasientdata <- pasientdata %>%
                  filter(!PID %in% sids_to_remove)

after_exclusion_count <- length(unique(pasientdata$PID))
# Update log
exclusion_log <- rbind(exclusion_log, data.frame(
  Step = 7,
  Reason = "Diagnosed before approval or after follow-up",
  ExcludedCount = before_exclusion_count - after_exclusion_count,
  RemainingCount = after_exclusion_count
))
```

7.4 Fjerner de med feil morfologi

```
before_exclusion_count <- length(unique(pasientdata$PID))
sids_to_remove <- c(1483,50,594)

# Remove records with the specified SIDs
pasientdata <- pasientdata %>%
                  filter(!PID %in% sids_to_remove)



after_exclusion_count <- length(unique(pasientdata$PID))
# Update log
exclusion_log <- rbind(exclusion_log, data.frame(
  Step = 8,
  Reason = "Correct morphology",
  ExcludedCount = before_exclusion_count - after_exclusion_count,
  RemainingCount = after_exclusion_count
))
```

7.5 Endrer verdinavn

```r
table(pasientdata$cTnmGruppe)
pasientdata <- pasientdata %>%
  mutate(cTnmGruppe = recode(cTnmGruppe,
                              "Ukjent" = "unknown",
                              ".m" = "missing"))

pasientdata <- pasientdata %>%
  mutate(funksjonsstatusUtr = recode(funksjonsstatusUtr,
                                     ".m" = "missing",
                                     "Ukjent" = "unknown",
                                     ".v" = "Not reported"))

pasientdata <- pasientdata %>%
  mutate(pdL1Resultat = recode(pdL1Resultat,
                               "0% eller negativ" = "0 or negative",
                               "Kan ikke vurderes" = "Cannot be assessed",
                               "Ikke angitt" = "Not specified",
                               .default = pdL1Resultat)) %>%
  mutate(pdL1Resultat = ifelse(pdL1Resultat == "" | pdL1Resultat == ".m", "Missing", pdL1Resultat))



pasientdata <- pasientdata %>%
  mutate(SEER_STADIUM = recode(SEER_STADIUM,
                               `1` = "Localized",
                               `2` = "Regional Metastasis",
                               `3` = "Distant Metastasis",
                               `9` = "Unknown"),
         SEER_STADIUM = factor(SEER_STADIUM, levels = c("Localized", "Regional Metastasis", "Distant Metastasis", "Unknown")))
```

```r
# Assuming pasientdata is your dataframe
pasientdata <- pasientdata %>%
  mutate(PERSONSTATUS = case_when(
        PERSONSTATUS == 2 ~ "Dead",
        PERSONSTATUS == 1 ~ "Alive",
        PERSONSTATUS == 3 ~ "Lost to follow-Up"
  ))

pasientdata <- pasientdata %>%
  mutate(KJOENN = recode(KJOENN,
                                 "M" = "Male",
                                 "K" = "Female"))




table(pasientdata$morfologigruppeLunge)
#sjekker orginale navn
original_unique_values <- unique(pasientdata$morfologigruppeLunge)



# Original unique values
original_unique_values <- c("Adenokarsinom", "IkkesmåcelletkarsinomUNS", "Plateepitelkarsinom")

# Name mapping including all original unique values
name_map <- setNames(
  c("Adenocarcinoma", "Non-small cell carcinoma NOS", "Squamous cell carcinoma"),
  original_unique_values
)

# Apply the name mapping to the 'morfologigruppeLunge' column in pasientdata
pasientdata$morfologigruppeLunge <- name_map[pasientdata$morfologigruppeLunge]
```

```
# Replace NA values in 'morfologigruppeLunge' with "Non-small cell carcinoma NOS"
pasientdata$morfologigruppeLunge[is.na(pasientdata$morfologigruppeLunge)] <- "Non-small cell carcinoma NOS"

# Check the updated table
table(pasientdata$morfologigruppeLunge)







table(pasientdata$funksjonsstatusUtr)

unique_values<- unique(pasientdata$SEER_STADIUM)
print(unique_values)
str(pasientdata)
```

7.6 legger inn straalebehandling

```
# Creating a simplified version of straaledata with just the PIDs
straaledata_simplified <- straaledata %>%
  select(PID) %>%
  distinct() %>%
  mutate(receivedRadiation = "Yes")

# Joining the data
pasientdata <- pasientdata %>%
  left_join(straaledata_simplified, by = "PID") %>%
  mutate(receivedRadiation = if_else(is.na(receivedRadiation), "No", receivedRadiation))
table(pasientdata$receivedRadiation)
```

8 eksporterer

```
write.csv(pasientdata, "noinputadherent.csv", row.names = FALSE)
table(pasientdata$cTnmGruppe)
```

8.1 histogram

```
# Aggregate data to count diagnoses per month
diagnosis_counts_per_month <- pasientdata %>%
  mutate(Month = floor_date(DIAGNOSEDATO, "month")) %>%
  group_by(Month) %>%
  summarize(Count = n())

# Create a histogram with a non-linear regression curve
histogram_with_curve <- ggplot(diagnosis_counts_per_month, aes(x = Month, y = Count)) +
  geom_col(fill = "blue") +  # Create histogram bars
  geom_smooth(method = "loess", se = FALSE, color = "red", span = 0.5) +  # Non-linear regression
  labs(x = "Month of Diagnosis",
       y = "Number of diagnosis in group 1",
       title = "Histogram of Diagnoses Per Month with LOESS trendline in group 1") +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "white"),
        plot.background = element_rect(fill = "white", color = "white"))  # Set entire plot background to white

# Save the plot
ggsave("diagnosis_histogram_with_curve.png", plot = histogram_with_curve, width = 10, height = 6, dpi = 300)
```

8.2 Kakediagram

```
# Categorize the time until first dose
pasientdata <- pasientdata %>%
              mutate(time_category = case_when(
                time_until_first_dose >= 0 & time_until_first_dose <= 92 ~ "within 1-3 months",
                time_until_first_dose >= 93 & time_until_first_dose <= 156 ~ "within 3-5 months"
,
            time_until_first_dose >= 157 & time_until_first_dose <= 366 ~ "5+ months",
            time_until_first_dose > 367 ~ "Over 12 months",  # Modified condition
            TRUE ~ "Unknown" # for any missing or out-of-range values
          ))

# Calculate count and percentage
time_summary <- pasientdata %>%
           group_by(cTnmGruppe, time_category) %>%
           summarise(count = n(), .groups = 'drop') %>%
           mutate(total = sum(count),
                percent = count / total)

# List of unique cTnmGruppe values
ctnm_values <- unique(time_summary$cTnmGruppe)

# Create a list to store plots
plots <- list()

# Define a softer color palette for the time categories
time_colors <- c("within 1-3 months" = "#add8e6",  # Light blue
            "within 3-5 months" = "#90ee90",  # Light green
            "5+ months" = "#ffcccb",      # Light red
            "Over 12 months" = "#ffd700",    # Gold
            "Unknown" = "#d3d3d3")          # Light grey

# The rest of your code remains the same until the plotting loop

for(ctnm in ctnm_values) {
  data_subset <- time_summary[time_summary$cTnmGruppe == ctnm,]
  p <- ggplot(data_subset, aes(x = "", y = percent, fill = time_category)) +
     geom_bar(stat = "identity", width = 1) +
     coord_polar("y", start = 0) +
     scale_fill_manual(values = time_colors) +
    theme_void() +
    labs(fill = "Time Category", title = paste("Cancer Stage:", ctnm)) +
    geom_text(aes(label = ifelse(percent > 0.001, scales::percent(percent), "")),
              position = position_stack(vjust = 0.5),
              size = 1.7,
              color = "black")
  plots[[ctnm]] <- p
}

# Combine your plots
combined_plot <- do.call(grid.arrange, c(plots, ncol = 2))

# Add an overall title
combined_plot_with_title <- arrangeGrob(combined_plot, top = "Time from diagnosis Until
 First Dose in group 1")

# Save the plot with the title
ggsave("diagnosis_firstdose_pie.png", combined_plot_with_title, width = 20, height = 20, uni
ts = "cm", dpi = 500)
```

8.3 Lager dag/mnd variabel

8.4 ny summary

```
#table3 <- pasientdata %>%
#  dplyr::select(
 #   funksjonsstatusUtr, morfologigruppeLunge, SEER_STADIUM,
#    pdL1Resultat, cTnmGruppe, ALDER, PERSONSTATUS, SurvivalTime, KJOENN, receive
dRadiation
#  ) %>%
#  rename(
 #   "ECOG status" = funksjonsstatusUtr,
  ## "SEER Stage" = SEER_STADIUM,
   # "PD-L1 precentage" = pdL1Resultat,
  #  "TNM Group" = cTnmGruppe,
  #  "Age (years)" = ALDER,
  # "Person Status" = PERSONSTATUS,
 #  "Gender" = KJOENN,
  # "Radiation therapy recieved" = receivedRadiation,
#    "Survival since diagnosis (days)" = SurvivalTime  ) %>%
#  tbl_summary(
 #   by = "TNM Group", # Stratify by TNM group
#    missing = "ifany" # Exclude missing data from the summary
 # ) %>%
#  add_overall() %>%
#  add_n() %>%
 # modify_header(label ~ "**Overview of all cases by cancer type**") %>%
 # bold_labels() %>%
 # modify_footnote(
  #  all_stat_cols() ~ "Values: n (%); Median (IQR). Note: Survival time is calculated up to October 2023. Data beyond this date are not available, which may limit the interpretation of long-term survival trends."
#  ) %>%

#  modify_caption("Table 2: Summary of Key Variables in the Whole Group")
```

##eksporterer

```
#gt_table <- as_gt(table3)
#gt::gtsave(gt_table, filename = "table3.png")
```

##kaplan meier uten 90 dagers filter

```
censoring_date <- as.Date("2022-12-31")

# Calculate Survival Time
pasientdata$survival_time <- ifelse(is.na(pasientdata$STATUSDATO),
                              as.numeric(difftime(censoring_date, pasientdata$first_dose, units = "days")),
                              as.numeric(difftime(pasientdata$STATUSDATO, pasientdata$first_dose, units = "days")))
```

```r
# Convert Survival Time to months and round to nearest month
pasientdata$survival_time_months <- round(pasientdata$survival_time / 30.44)

# Censoring status: 0 if alive (censored), 1 if dead (event occurred)
pasientdata$censoring_status <- ifelse(is.na(pasientdata$STATUSDATO), 0, 1)

# Create the survival object using months
surv_obj <- Surv(time = pasientdata$survival_time_months, event = pasientdata$censoring_status)

# Fit the Kaplan-Meier survival curve
km_fit <- survfit(surv_obj ~ 1, data = pasientdata)

# Calculate survival probabilities at specific time points (36, 48, and 60 months)
time_points_months <- c(36, 48, 60)  # 3, 4, and 5 years in months
survival_probabilities <- summary(km_fit, times = time_points_months)$surv

# Extract survival probabilities
survival_at_3_years <- survival_probabilities[1]
survival_at_4_years <- survival_probabilities[2]
survival_at_5_years <- survival_probabilities[3]

# Find the median survival time in months
median_survival_months <- summary(km_fit)$table['median']

# Extract the survival curve data
surv_data <- broom::tidy(km_fit, conf.int = TRUE)

# Calculate differences in months from the median to the lower and upper CI bounds
lower_ci_diff <- surv_data$time[which.max(surv_data$conf.low <= 0.5)]
upper_ci_diff <- surv_data$time[which.max(surv_data$conf.high <= 0.5)]

# Calculate the CI for the median survival time
```

```r
lower_ci_median <- lower_ci_diff
upper_ci_median <- upper_ci_diff

survival_plot <- ggsurvplot(
  km_fit, data = pasientdata, conf.int = TRUE,
  risk.table = TRUE,
  xlab = "Months since first dose", ylab = "Survival probability",
  title = "Kaplan-Meier Survival Curve of group 1",
  ggtheme = theme_minimal(),
  break.x.by = 3,  # Set x-axis breaks every 10 months
   xlim = c(0, 70),  # Extend x-axis to 60 months
  risk.table.fontsize = 6  # Adjust the font size of the numbers at risk (12 in this example)
)

survival_plot$plot <- survival_plot$plot +
  theme(
    plot.title = element_text(size = 19),  # Adjust the size of the title
    axis.title = element_text(size = 15),
    axis.text = element_text(size = 15))  # Adjust the size of the numbers at risk

# Add median survival line and year markers
year_markers <- c(36, 48, 60)  # 3, 4, and 5 years in months
survival_plot$plot <- survival_plot$plot +
  geom_vline(xintercept = year_markers, linetype = "dotted", color = "black") +
  geom_vline(xintercept = median_survival_months, linetype = "dashed", color = "black")

# Offset for annotations
offset <- -5.7
offset2 <- -9.7
# Annotate the median survival time and its CI
median_annotation <- paste("Median:", round(median_survival_months, 1), "months\nCI:",
                          round(lower_ci_median, 1), "-", round(upper_ci_median, 1), "months")
survival_plot$plot <- survival_plot$plot +
  annotate("text", x = median_survival_months - offset2, y = 0.8, label = median_annotation,
size = 6, vjust = -0.5, color = "black") +
  annotate("text", x = 36 - offset, size = 6, y = 0.7, label = paste(round(survival_at_3_years * 100, 1), "% at 3 yrs"), color = "black") +
  annotate("text", x = 48 - offset, size = 6, y = 0.6, label = paste(round(survival_at_4_years * 100, 1), "% at 4 yrs"), color = "black") +
  annotate("text", x = 60 - offset, size = 6, y = 0.5, label = paste(round(survival_at_5_years * 100, 1), "% at 5 yrs"), color = "black")

# Print and save the plot
print(survival_plot)

# Combine the plot and risk table into a single grid object
km_combined_grid <- arrangeGrob(survival_plot$plot, survival_plot$table, ncol = 1, heights = c(5, 1))



# Save the combined plot as an image
ggsave("avvikKMgruppe1included.png", km_combined_grid, width = 15, height = 11, bg = "#FFFDFB", dpi = 300)
```

##kaplan meier ecog 2+

```
pasientdata2 <- pasientdata[pasientdata$funksjonsstatusUtr >= 2, ]


pasientdata2 <- pasientdata2 %>%
  filter(pasientdata2$funksjonsstatusUtr != "missing")
pasientdata2 <- pasientdata2 %>%
  filter(pasientdata2$funksjonsstatusUtr != "unknown")
pasientdata2 <- pasientdata2 %>%
  filter(pasientdata2$funksjonsstatusUtr != "Not reported")
pasientdata2 <- pasientdata2

table (pasientdata2$funksjonsstatusUtr)
table (pasientdata2$morfologigruppeLunge)
mean(pasientdata2$ALDER)
```

```r
# Create the survival object using months
surv_obj <- Surv(time = pasientdata2$survival_time_months, event = pasientdata2$censoring_status)

# Fit the Kaplan-Meier survival curve
km_fit <- survfit(surv_obj ~ 1, data = pasientdata2)

# Calculate survival probabilities at specific time points (36, 48, and 60 months)
time_points_months <- c(36, 48, 60)  # 3, 4, and 5 years in months
survival_probabilities <- summary(km_fit, times = time_points_months)$surv

# Extract survival probabilities
survival_at_3_years <- survival_probabilities[1]
survival_at_4_years <- survival_probabilities[2]
survival_at_5_years <- survival_probabilities[3]

# Find the median survival time in months
median_survival_months <- summary(km_fit)$table['median']

# Extract the survival curve data
surv_data <- broom::tidy(km_fit, conf.int = TRUE)

# Calculate differences in months from the median to the lower and upper CI bounds
lower_ci_diff <- surv_data$time[which.max(surv_data$conf.low <= 0.5)]
upper_ci_diff <- surv_data$time[which.max(surv_data$conf.high <= 0.5)]

# Calculate the CI for the median survival time
lower_ci_median <- lower_ci_diff
upper_ci_median <- upper_ci_diff

survival_plot <- ggsurvplot(
  km_fit, data = pasientdata2, conf.int = TRUE,
  risk.table = TRUE,
  xlab = "Months since first dose", ylab = "Survival probability",
  title = "Kaplan-Meier Survival Curve of patients with ECOG status >1",
  ggtheme = theme_minimal(),
  break.x.by = 3,  # Set x-axis breaks every 10 months
  xlim = c(0, 70),  # Extend x-axis to 60 months
  risk.table.fontsize = 6  # Adjust the font size of the numbers at risk (12 in this example)
)

survival_plot$plot <- survival_plot$plot +
  theme(
    plot.title = element_text(size = 19),  # Adjust the size of the title
    axis.title = element_text(size = 15),
    axis.text = element_text(size = 15))  # Adjust the size of the numbers at risk

# Add median survival line and year markers
year_markers <- c(36, 48, 60)  # 3, 4, and 5 years in months
survival_plot$plot <- survival_plot$plot +
  geom_vline(xintercept = year_markers, linetype = "dotted", color = "black") +
  geom_vline(xintercept = median_survival_months, linetype = "dashed", color = "black")

# Offset for annotations
offset <- -5.7
offset2 <- -9.7
# Annotate the median survival time and its CI
median_annotation <- paste("Median:", round(median_survival_months, 1), "months\nCI:",
                      round(lower_ci_median, 1), "-", round(upper_ci_median, 1), "months")
survival_plot$plot <- survival_plot$plot +
  annotate("text", x = median_survival_months - offset2, y = 0.8, label = median_annotation,
size = 6, vjust = -0.5, color = "black") +
  annotate("text", x = 36 - offset, size = 6, y = 0.7, label = paste(round(survival_at_3_years
* 100, 1), "% at 3 yrs"), color = "black") +
  annotate("text", x = 48 - offset, size = 6, y = 0.6, label = paste(round(survival_at_4_years
* 100, 1), "% at 4 yrs"), color = "black") +
  annotate("text", x = 60 - offset, size = 6, y = 0.5, label = paste(round(survival_at_5_years
* 100, 1), "% at 5 yrs"), color = "black")

# Print and save the plot


print(survival_plot)

# Combine the plot and risk table into a single grid object
km_combined_grid <- arrangeGrob(survival_plot$plot, survival_plot$table, ncol = 1, heights = c(5, 1))


# Save the combined plot as an image
ggsave("avvikKMgruppe1included222.png", km_combined_grid, width = 15, height = 11, bg = "#FFFDFB", dpi = 300)
```

#lagrer csv fil

```
write.csv(pasientdata, file = "rwdpasientdatamedavvik.csv", row.names = FALSE)
```

9 cox regresjon

```
pasientdatacox<- pasientdata
table(pasientdata$funksjonsstatusUtr)
table(pasientdatacox$cTnmGruppe)

## Filtrerer ut alle som ikke er stadie 4. mistenker at gamle tall på variabler som pdl1 kan påvirke koeffisienter


##sjekk med funksjonstatus
table(pasientdatacox$funksjonsstatusUtr)

#fjerner han med ecog 4 pga lav sample size
pasientdatacox <- pasientdatacox %>%
  filter(PID != 1951)
```

```
pasientdatacox <- pasientdatacox %>%
  mutate(funksjonsstatusUtr = factor(case_when(
    funksjonsstatusUtr %in% c("Not reported", "unknown", "missing") ~ NA_character_,
    TRUE ~ funksjonsstatusUtr
  )))


pasientdatacox$pdL1Resultat <- as.character(pasientdatacox$pdL1Resultat)
pasientdatacox$pdL1Resultat[pasientdatacox$pdL1Resultat %in% c("Cannot be assessed", "
Missing", "Not specified")] <- NA
pasientdatacox$pdL1Resultat[pasientdatacox$pdL1Resultat == "<1"] <- "0 or negative"


##SJEKKER ALDER DISTRUBISJON
pasientdatacox %>%
  ggplot(aes(x = ALDER)) +
  geom_histogram()

pasientdatacox$PdL1Resultat <- factor(pasientdatacox$pdL1Resultat)



cox_model <- coxph(Surv(survival_time_months, censoring_status) ~ KJOENN + ALDER +
 funksjonsstatusUtr + pdL1Resultat, data = pasientdatacox)


test_ph <- cox.zph(cox_model)
print(test_ph) # Viser testresultatene
plot(test_ph)  # Plotter Schoenfeld residuals

# View the summary of the model
summary(cox_model)
table(pasientdatacox$funksjonsstatusUtr)
```

9.1 senstest ved og ecog 0-1

```
##bruk pasientdata cox dersom du ønsker å filtrere for stadie 4 også
pasientdata_filtered <- pasientdata[pasientdata$funksjonsstatusUtr <= 1, ]
table(pasientdata$funksjonsstatusUtr)
table(pasientdata_filtered$funksjonsstatusUtr)
table(pasientdata_filtered$cTnmGruppe)

# Assuming you have already created pasientdata_filtered as per previous steps

# Define the censoring date (end of 2022)
censoring_date <- as.Date("2022-12-31")

# Calculate Survival Time for the filtered data
pasientdata_filtered$survival_time <- ifelse(is.na(pasientdata_filtered$STATUSDATO),
                                        as.numeric(difftime(censoring_date, pasientdata_filtered$first_dose, units = "days")),
                                        as.numeric(difftime(pasientdata_filtered$STATUSDATO, pasientdata_filtered$first_dose, units = "days")))

# Convert Survival Time to months and round to nearest month
pasientdata_filtered$survival_time_months <- round(pasientdata_filtered$survival_time / 30.44)

# Censoring status: 0 if alive (censored), 1 if dead (event occurred)
pasientdata_filtered$censoring_status <- ifelse(is.na(pasientdata_filtered$STATUSDATO), 0, 1)

# Create the survival object using months for the filtered data
surv_obj <- Surv(time = pasientdata_filtered$survival_time_months, event = pasientdata_filtered$censoring_status)

# Fit the Kaplan-Meier survival curve for the filtered data
km_fit <- survfit(surv_obj ~ 1, data = pasientdata_filtered)
```

```r
# Calculate survival probabilities at specific time points (36, 48, and 60 months) for the filtered data
time_points_months <- c(36, 48, 60)
survival_probabilities <- summary(km_fit, times = time_points_months)$surv

# Extract survival probabilities for the filtered data
survival_at_3_years <- survival_probabilities[1]
survival_at_4_years <- survival_probabilities[2]
survival_at_5_years <- survival_probabilities[3]

# Find the median survival time in months for the filtered data
median_survival_months <- summary(km_fit)$table['median']

# Extract the survival curve data for the filtered data
surv_data <- broom::tidy(km_fit, conf.int = TRUE)

# Calculate differences in months from the median to the lower and upper CI bounds for the filtered data
lower_ci_diff <- surv_data$time[which.max(surv_data$conf.low <= 0.5)]
upper_ci_diff <- surv_data$time[which.max(surv_data$conf.high <= 0.5)]

# Calculate the CI for the median survival time for the filtered data
lower_ci_median <- lower_ci_diff
upper_ci_median <- upper_ci_diff

survival_plot <- ggsurvplot(
  km_fit, data = pasientdata_filtered, conf.int = TRUE,
  risk.table = TRUE,
  xlab = "Months since first dose", ylab = "Survival probability",
  title = "Kaplan-Meier Survival Curve excluding patients with ECOG status above 1",
  ggtheme = theme_minimal(),
 break.x.by = 3,  # Set x-axis breaks every 10 months
  xlim = c(0, 70)  # Extend x-axis to 60 months
)
```

```r
# Add median survival line and year markers
year_markers <- c(36, 48, 60)  # 3, 4, and 5 years in months
survival_plot$plot <- survival_plot$plot +
  geom_vline(xintercept = year_markers, linetype = "dotted", color = "black") +
  geom_vline(xintercept = median_survival_months, linetype = "dashed", color = "black")

# Offset for annotations
offset <- -5.5

# Annotate the median survival time and its CI
median_annotation <- paste("Median:", round(median_survival_months, 1), "months\nCI:",
                           round(lower_ci_median, 1), "-", round(upper_ci_median, 1), "months")
survival_plot$plot <- survival_plot$plot +
  annotate("text", x = median_survival_months - offset, y = 0.8, label = median_annotation, vjust = -0.5, color = "black") +
  annotate("text", x = 36 - offset, y = 0.7, label = paste(round(survival_at_3_years * 100, 1), "% at 3 yrs"), color = "black") +
  annotate("text", x = 48 - offset, y = 0.6, label = paste(round(survival_at_4_years * 100, 1), "% at 4 yrs"), color = "black") +
  annotate("text", x = 60 - offset, y = 0.5, label = paste(round(survival_at_5_years * 100, 1), "% at 5 yrs"), color = "black")



print(survival_plot)

# Combine the plot and risk table into a single grid object
km_combined_grid <- arrangeGrob(survival_plot$plot, survival_plot$table, ncol = 1, heights = c(5, 1))



# Save the combined plot as an image
ggsave("avvikKMgruppe1Ecog.png", km_combined_grid, width = 15, height = 11, bg = "#FFFDFB", dpi = 300)
```

```
table(pasientdata_filtered$time_category)
```

9.2 stratifisering 2019

```
censoring_date <- as.Date("2022-12-31")

# Calculate Survival Time
pasientdata$survival_time <- ifelse(is.na(pasientdata$STATUSDATO),
                                    as.numeric(difftime(censoring_date, pasientdata$first_dose, units = "days")),
                                    as.numeric(difftime(pasientdata$STATUSDATO, pasientdata$first_dose, units = "days")))

# Convert Survival Time to months and round to nearest month
pasientdata$survival_time_months <- round(pasientdata$survival_time / 30.44)

# Censoring status: 0 if alive (censored), 1 if dead (event occurred)
pasientdata$censoring_status <- ifelse(is.na(pasientdata$STATUSDATO), 0, 1)

# Create the survival object using months
surv_obj <- Surv(time = pasientdata$survival_time_months, event = pasientdata$censoring_status)
# Create a new column for stratification
pasientdata$pre_post_2019 <- ifelse(year(pasientdata$first_dose) < 2019, "Pre-2019", "Post-2019")
# Create the survival object using months with stratification
surv_obj_stratified <- Surv(time = pasientdata$survival_time_months, event = pasientdata$censoring_status)

# Fit the Kaplan-Meier survival curve with stratification
km_fit_stratified <- survfit(surv_obj_stratified ~ pre_post_2019, data = pasientdata)

# Generate the survival plot with stratification
survival_plot_stratified <- ggsurvplot(
  km_fit_stratified, data = pasientdata, conf.int = TRUE,
  risk.table = TRUE,
  xlab = "Months since first dose", ylab = "Survival probability",
  title = "Kaplan-Meier Survival Curve by Pre and Post 2019",
  ggtheme = theme_minimal(),
  break.x.by = 3  # Set x-axis breaks every 10 months
)

# Print and save the stratified plot
print(survival_plot_stratified)

# Perform the log-rank test
log_rank_test <- survdiff(surv_obj_stratified ~ pre_post_2019, data = pasientdata)


# Print the results of the log-rank test
print(log_rank_test)
```

9.3 med filter

```
before_exclusion_count <- length(unique(pasientdata$PID))




# Define the censoring date (end of 2022)
censoring_date <- as.Date("2022-12-31")

# Calculate Survival Time
pasientdata$survival_time <- ifelse(is.na(pasientdata$STATUSDATO),
                                     as.numeric(difftime(censoring_date, pasientdata$first_dose, units = "days")),
                                     as.numeric(difftime(pasientdata$STATUSDATO, pasientdata$first_dose, units = "days")))

# Convert Survival Time to months and round to nearest month
pasientdata$survival_time_months <- round(pasientdata$survival_time / 30.44)
```

```r
# Exclude negative or very short survival times (less than 3 months)
pasientdata <- pasientdata %>%
        filter(survival_time_months >= 1)

# Censoring status: 0 if alive (censored), 1 if dead (event occurred)
pasientdata$censoring_status <- ifelse(is.na(pasientdata$STATUSDATO), 0, 1)

# Create the survival object using months
surv_obj <- Surv(time = pasientdata$survival_time_months, event = pasientdata$censoring_status)

# Fit the Kaplan-Meier survival curve
km_fit <- survfit(surv_obj ~ 1, data = pasientdata)

# Calculate survival probabilities at specific time points (36, 48, and 60 months)
time_points_months <- c(36, 48, 60)  # 3, 4, and 5 years in months
survival_probabilities <- summary(km_fit, times = time_points_months)$surv

# Extract survival probabilities
survival_at_3_years <- survival_probabilities[1]
survival_at_4_years <- survival_probabilities[2]
survival_at_5_years <- survival_probabilities[3]

# Find the median survival time in months
median_survival_months <- summary(km_fit)$table['median']

# Extract the survival curve data
surv_data <- broom::tidy(km_fit, conf.int = TRUE)

# Calculate differences in months from the median to the lower and upper CI bounds
lower_ci_diff <- surv_data$time[which.max(surv_data$conf.low <= 0.5)]
upper_ci_diff <- surv_data$time[which.max(surv_data$conf.high <= 0.5)]
```

```r
# Calculate the CI for the median survival time
lower_ci_median <- lower_ci_diff
upper_ci_median <- upper_ci_diff

survival_plot <- ggsurvplot(
  km_fit, data = pasientdata, conf.int = TRUE,
  risk.table = TRUE,
  xlab = "Months since first dose", ylab = "Survival probability",
  title = "Kaplan-Meier Survival Curve excluding death within 1 month in group 1",
  ggtheme = theme_minimal(),
 break.x.by = 3,  # Set x-axis breaks every 10 months
  xlim = c(0, 70)  # Extend x-axis to 60 months
)

# Add median survival line and year markers
year_markers <- c(36, 48, 60)  # 3, 4, and 5 years in months
survival_plot$plot <- survival_plot$plot +
  geom_vline(xintercept = year_markers, linetype = "dotted", color = "black") +
  geom_vline(xintercept = median_survival_months, linetype = "dashed", color = "black")

# Offset for annotations
offset <- -5.5

# Annotate the median survival time and its CI
median_annotation <- paste("Median:", round(median_survival_months, 1), "months\nCI:",
                           round(lower_ci_median, 1), "-", round(upper_ci_median, 1), "months")
survival_plot$plot <- survival_plot$plot +
  annotate("text", x = median_survival_months - offset, y = 0.8, label = median_annotation, vjust = -0.5, color = "black") +
  annotate("text", x = 36 - offset, y = 0.7, label = paste(round(survival_at_3_years * 100, 1), "% at 3 yrs"), color = "black") +
  annotate("text", x = 48 - offset, y = 0.6, label = paste(round(survival_at_4_years * 100, 1), "% at 4 yrs"), color = "black") +
  annotate("text", x = 60 - offset, y = 0.5, label = paste(round(survival_at_5_years * 100, 1),
"% at 5 yrs"), color = "black")

# Print and save the plot
print(survival_plot)

# Combine the plot and risk table into a single grid object
km_combined_grid <- arrangeGrob(survival_plot$plot, survival_plot$table, ncol = 1, heights = c(5, 1))

# Save the combined plot as an image
ggsave("avvikKMGruppe1excluded.png", km_combined_grid, width = 15, height = 11, bg = "#FFFDFB", dpi = 300)





after_exclusion_count <- length(unique(pasientdata$PID))

# Update log
exclusion_log <- rbind(exclusion_log, data.frame(
  Step = 9,
  Reason = "Survival 1 months after treatment initiation",
  ExcludedCount = before_exclusion_count - after_exclusion_count,
  RemainingCount = after_exclusion_count
))
```

9.4 eksporterer exclusion logg

```r
write.csv(exclusion_log, "exclusion_logGruppe1.csv", row.names = FALSE)

pasientdatacoxdeath <- pasientdata[pasientdata$funksjonsstatusUtr <= 1, ]

write.csv(pasientdata_filtered, file = "rwdecog.csv", row.names = FALSE)
write.csv(pasientdatacoxdeath, file = "rwdcoxdeath.csv", row.names = FALSE)
```

```r
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(echo = TRUE)
# Detect the computer's name
computer_name <- Sys.info()["nodename"]

# Set the working directory based on the computer's name
if (computer_name == "JEROEN-LAPTOP") {
  setwd("C:\\masteroppgaven lokal\\raw data\\0 inputtering")
} else if (computer_name == "JEROENHAUKAAS") {
  setwd("D:/Masteroppgaven backup/raw data/0 inputtering")
} else {
  stop("Unknown computer: unable to set the working directory")
}

# Load data
inputertpdl1 <- read.csv("Utlevert_kreftdata_4082.csv", header = TRUE, sep = ";")
datasykehus <- read.csv("Utlevert_mkb_sykehus_4082.CSV", header = TRUE, sep = ";")

inputertpdl1$STATUSDATO <- as.Date(inputertpdl1$STATUSDATO, format = '%d%b%Y'
)

exclusion_log <- data.frame(
  Step = integer(),
  Reason = character(),
  ExcludedCount = integer(),
  RemainingCount = integer(),
  stringsAsFactors = FALSE
)

before_exclusion_count <- length(unique(inputertpdl1$PID))
after_exclusion_count <- length(unique(inputertpdl1$PID))

# Update log
exclusion_log <- rbind(exclusion_log, data.frame(
  Step = 1,
  Reason = "Initial patients",
  ExcludedCount = before_exclusion_count - after_exclusion_count,
  RemainingCount = after_exclusion_count
))



library(tidyverse)
library(ggplot2)
library(janitor)
library(mice)
library(dplyr)
library(DataExplorer)
library(lubridate)
```

10 Oversikt variabler

```
# lager rapport med alle variablene, sjekk fillokasjonen RMD filen er lagret. Kan også brukes
 på individuelle variabler
#create_report(datasykehus)
summary(datasykehus)
names(datasykehus)

# Assuming your dataframe is datasykehus, and it has columns PID and behregime

# Removing duplicates: Keeping only the first occurrence of each PID within each behregime
unique_datasykehus <- datasykehus %>%
  group_by(behregime) %>%
  distinct(PID, .keep_all = TRUE)

# Now create the frequency table for behregime based on unique PIDs
table(unique_datasykehus$behregime)
```

##Endrer datoformat

```
# endrer datoformat datoAdministrasjonVirkestoff
datasykehus$datoAdministrasjonVirkestoff <- as.Date(datasykehus$datoAdministrasjonVirkestoff, format = '%d%b%Y')
```

##Legger inn diagnosedato basert på PID

```
# Legger inn diagnosedato basert på PID
datasykehusdiag <- datasykehus %>%
  left_join(inputertpdl1[, c("PID", "DIAGNOSEDATO")], by = "PID")
str(datasykehusdiag)
#diagnosedato er definert som bokstav, må endres
datasykehusdiag$DIAGNOSEDATO <- as.Date(datasykehusdiag$DIAGNOSEDATO, format = "%d%b%Y")
#sjekk
str(datasykehusdiag)
```

##Henter ut alle med pembro behandling Hele denne kan hoppes over.

```
# Ekstraherer alle pembrolizumab behandlinger
allekeytruda<- datasykehusdiag[grepl("pembrolizumab", datasykehusdiag$behregime, ignore.case = TRUE), ]

# sjekk
table(allekeytruda$behregime)
antallkeytruda <- length(unique(allekeytruda$PID))

# Henter ut de som KUN har motatt monoterapi (aldri kombo eller annen behandling) i løpet av hele oppfølgingstiden
pembrolizumab_mono_patients <- datasykehusdiag %>%
  group_by(PID) %>%
  filter(all(behregime == "Pembrolizumab")) %>%
  ungroup()
#sjekker hvor mange av de som kun har fått monoterapi
pembrolizumabmono<- pembrolizumab_mono_patients %>%
  filter(behregime == "Pembrolizumab")
```

```
# antall personer kun mono
unique_pids1 <- n_distinct(pembrolizumab_mono_patients$PID)
cat("Number of unique PIDs: ", unique_pids1, "\n")
```

##Fjerner de som ikke mottar pembrolizumab mono i det heletatt.

Vi går fra 2970 som har motatt noen form for pembro til 1927 som har motatt monoterapi i en viss periode.

```
#denne koden inkluderer alle som har fått pembrro innen x måneder fra diagnosedatoen, men
 de kan også ha fått andre behandlinger senere
datasykehus3<- allekeytruda %>%
  filter(behregime == "Pembrolizumab")
# Check the resulting data
unique_pids2 <- n_distinct(datasykehus3$PID)
cat("Number of unique PIDs: ", unique_pids2, "\n")
```

10.1 Henter andre behandlinger tilbake

```
before_exclusion_count <- length(unique(inputertpdl1$PID))
##henter tilbake annen data på de som har brukt monoterapi innen 1 mnd.
datasykehus4 <- datasykehus %>%
  filter(PID %in% datasykehus3$PID)

unique_pids3 <- n_distinct(datasykehus4$PID)
cat("Number of unique PIDs: ", unique_pids3, "\n")
table(datasykehus4$behregime)

after_exclusion_count <- length(unique(datasykehus4$PID))

exclusion_log <- rbind(exclusion_log, data.frame(
  Step = 2,
  Reason = "Recieved pembrolizumab monotherapy",
  ExcludedCount = before_exclusion_count - after_exclusion_count,
  RemainingCount = after_exclusion_count
))
```

##sorterer basert på dato legemiddelbruk

```
datasykehus5 <- datasykehus4 %>%
  arrange(PID, datoAdministrasjonVirkestoff)
table(datasykehus5$behregime)
```

10.2 lager liste for oppstartsdato for pasientene

```
datasykehus5$datoAdministrasjonVirkestoff <- as.Date(datasykehus5$datoAdministrasjonVirkestoff, format = "%d%b%Y")

# startdato settes som første dose med pembro, så kombo tells også



startpembro <- datasykehus5 %>%
  filter(behregime == "Pembrolizumab") %>%
  group_by(PID) %>%
  summarise(first_dose = min(datoAdministrasjonVirkestoff)) %>%
  ungroup()

# Step 2: Merge this information back into the original dataset.
datasykehus6 <- datasykehus5 %>%
  left_join(startpembro, by = "PID")
```

10.3 Setter inn sluttdato

bruker 23 måneder for sikkerhetskyld. behandlingen skal ta 105 uker.

```
datasykehus6$first_dose <- as.Date(datasykehus6$first_dose, format = "%d%b%Y")

datasykehus7 <- datasykehus6 %>%
  mutate(
    last_dose = first_dose %m+% months(24))
```

10.4 flagger ned alle som har fått annen behandling før behandlingstiden var over

```
##Lager en variabel som sier om noe er et avvik fra behandlingsregimet. Definerer avvik som behandlingsregime som ikke er pembro innen 2 aar fra forste dose.

#Lager en variabel til som sier om vedkommende avviker i løpet av behandlingen
datasykehus8 <- datasykehus7 %>%
  mutate(
    IsDeviation = ifelse(behregime != "Pembrolizumab" &
                         datoAdministrasjonVirkestoff <= last_dose &
                         datoAdministrasjonVirkestoff >= first_dose,
                         TRUE,
                         FALSE)
  ) %>%
  group_by(PID) %>%
  # Determine if there was any deviation for each patient
  mutate(
    PatientDeviation = any(IsDeviation)
  ) %>%
  ungroup()

pre_pembro_treatments <- datasykehus7 %>%
  filter(datoAdministrasjonVirkestoff < first_dose) %>%
  group_by(PID) %>%
  arrange(PID, datoAdministrasjonVirkestoff) %>%
  summarize(TreatmentsBeforePembro = list(behregime))
print(pre_pembro_treatments)



before_exclusion_count <- length(unique(datasykehus7$PID))


#fjerner de som har motatt monoklonale antistoffer
datasykehus8 <- datasykehus7 %>%
  mutate(
    IsDeviation = ifelse(behregime != "Pembrolizumab" &
                         datoAdministrasjonVirkestoff <= last_dose &
                         datoAdministrasjonVirkestoff >= first_dose,
                         TRUE,
                         FALSE),
    ThreeMonthsBeforeFirstDose = first_dose %m-% months(3),
    TwelveMonthsBeforeFirstDose = first_dose %m-% months(12)
  ) %>%
  group_by(PID) %>%
  mutate(
    PatientDeviation = any(IsDeviation),
    ReceivedAntistoffer = any((typeMkb == "Monoklonale antistoffer") &
                              (datoAdministrasjonVirkestoff < first_dose) &
                              (datoAdministrasjonVirkestoff >= TwelveMonthsBeforeFirstDose) &
                              (datoAdministrasjonVirkestoff < ThreeMonthsBeforeFirstDose))
  ) %>%
  ungroup() %>%
  # Filter out patients who received Monoklonale antistoffer 3-12 months before first dose
  filter(!ReceivedAntistoffer)

after_exclusion_count <- length(unique(datasykehus8$PID))
exclusion_log <- rbind(exclusion_log, data.frame(
  Step = 3,
  Reason = "No recent cancer treatment with immunotherapies",
  ExcludedCount = before_exclusion_count - after_exclusion_count,
  RemainingCount = after_exclusion_count
))
```

10.5 Deler opp pasientene i de som fullfører kuren og de som ikke gjør

```
# Patients with no deviations
before_exclusion_count <- length(unique(datasykehus8$PID))


no_deviation_patients <- datasykehus8 %>%
  filter(PatientDeviation == FALSE)

#teller
unique_pids4 <- n_distinct(no_deviation_patients$PID)
cat("Number of unique PIDs: ", unique_pids4, "\n")



# Patients with deviations
deviation_patients <- datasykehus8 %>%
  filter(PatientDeviation == TRUE)

after_exclusion_count <- length(unique(no_deviation_patients$PID))
```

#undersøkelser ## Undersøker totaldosen pembro (mg) til pasientene som fullførte kuren

```
# legger inn dødsdato
fullfort <- no_deviation_patients %>%
  left_join(inputertpdl1[, c("PID", "STATUSDATO")], by = "PID")
#STATUSDATO er definert som bokstav, må endres

fullfort$STATUSDATO <- as.Date(fullfort$STATUSDATO, format = "%d%b%Y")
str(fullfort)
fullfort <- fullfort %>%
  mutate(
    doseVirkestoff = ifelse(typeMkb == "Monoklonale antistoffer", doseVirkestoff, NA)
  )

#ser på de som har fullført kuren
cutoff_date <- as.Date("2023-01-1")

fullfort <- fullfort %>%
  filter(IsDeviation == FALSE) %>%
  filter(last_dose < cutoff_date) %>%
  filter(is.na(STATUSDATO) | STATUSDATO > last_dose)
#henter ut doser
```

```
total_dosage_by_patient <- fullfort %>%
  group_by(PID) %>%
  summarise(TotalDosage = sum(doseVirkestoff, na.rm = TRUE))


#Plotter det
ggplot(total_dosage_by_patient, aes(x = TotalDosage)) +
  geom_histogram(binwidth = 500, fill = "blue", color = "black") +
  labs(title = "Histogram of Total Pembrolizumab Dosage per Patient Who Survived Until the
End Date",
       x = "Total Dosage (mg)",
       y = "Count of Patients") +
  theme_minimal()
```

10.6 undersøker de med høy og lave dose

```
high_dosage_patients <- total_dosage_by_patient %>%
  filter(TotalDosage > 20000)
```

10.7 Forsøk på å telle antall folk og lage graf

10.8 teller antall som motokk ulike behandlingsregimer blant avvikende

```
#teller hvor mange som har motatt ulike behandlingsregimer
# Count unique PIDs for each treatment regimen
unique_pid_per_regimen <- deviation_patients %>%
  group_by(behregime) %>%
  summarise(UniquePIDCount = n_distinct(PID)) %>%
  arrange(desc(UniquePIDCount))

# View the result
print(unique_pid_per_regimen)
total_unique_pid <- deviation_patients %>%
  summarise(TotalUniquePIDCount = n_distinct(PID))
total_unique_pid2 <- allekeytruda %>%
  summarise(TotalUniquePIDCount = n_distinct(PID))
```

```
boxplot(no_deviation_patients$first_dose)
```

#eksporterer datasettene

```
# sjekk hvor dataen havner
getwd()
# endrer dato for første dose og ser hovordan det påvirker snittdato
datasykehus8$first_dose <- as.Date(datasykehus8$first_dose)

# Convert dates to numeric (days since a reference date, e.g., 1970-01-01)
numeric_dates <- as.numeric(datasykehus8$first_dose)

# Calculate average numeric date
average_numeric_date <- mean(numeric_dates, na.rm = TRUE)

#Justerer startdato til første pembro dose, ikke første datasykehus8dose

no_deviation_patients <- no_deviation_patients %>%
  filter(virkestoff == "Pembrolizumab") %>%
  group_by(PID) %>%
 summarise(first_dose = min(datoAdministrasjonVirkestoff)) %>%
  ungroup()

datasykehus8 <- datasykehus8 %>%
  filter(virkestoff == "Pembrolizumab") %>%
  group_by(PID) %>%
  summarise(first_dose = min(datoAdministrasjonVirkestoff)) %>%
  ungroup()

boxplot(no_deviation_patients$first_dose)

datasykehus8$first_dose <- as.Date(datasykehus8$first_dose)

# Convert dates to numeric (days since a reference date, e.g., 1970-01-01)
numeric_dates <- as.numeric(datasykehus8$first_dose)
```

~~~
average_date <- as.Date(average_numeric_date, origin = "1970-01-01")


# Calculate average numeric date
average_numeric_date2 <- mean(numeric_dates, na.rm = TRUE)
average_date2 <- as.Date(average_numeric_date2, origin = "1970-01-01")

write.csv(datasykehus8, file = "medavvik.csv", row.names = FALSE)
#skriver ut datasettet uten avvik
write.csv(no_deviation_patients, file = "monoterapiutenavvik.csv", row.names = FALSE)


#skriver ut datasettet uten avvik
write.csv(deviation_patients, file = "monoterapimedavvik.csv", row.names = FALSE)
write.csv(allekeytruda, file = "keytruda.csv", row.names = FALSE)


write.csv(exclusion_log, "exclusion_log.csv", row.names = FALSE)
~~~

~~~
```
~~~

~~~
knitr::opts_chunk$set(echo = TRUE)
# Detect the computer's name
computer_name <- Sys.info()["nodename"]


# Set the working directory based on the computer's name
if (computer_name == "JEROEN-LAPTOP") {
  setwd("C:\\masteroppgaven lokal\\raw data\\0 inputtering")
} else if (computer_name == "JEROENHAUKAAS") {
  setwd("D:/Masteroppgaven backup/raw data/0 inputtering")
} else {
  stop("Unknown computer: unable to set the working directory")
}



# Load data
datahresept <- read.csv("Utlevert_mkb_hresept_4082.csv", header = TRUE, sep = ";")
monoterapi <- read.csv("medavvik.csv", header = TRUE, sep = ",")
~~~

```r
inputertpdl1 <- read.csv("Utlevert_kreftdata_4082.csv", header = TRUE, sep = ";")
exclusion_log <- read.csv("exclusion_log.csv")



library(tidyverse)
library(ggplot2)
library(janitor)
library(mice)
library(dplyr)
library(DataExplorer)
library(lubridate)
```

11 Oversikt variabler utlevert kreftdata kan hoppes over til ## rens av variabler

```r
# lager rapport med alle variablene, sjekk fillokasjonen RMD filen er lagret. Kan også brukes på individuelle variabler
#create_report(datahresept )
#summary(datahresept)
#names(datahresept)
```

#Rens ##endrer datoformat

```r
monoterapi$first_dose <- as.Date(monoterapi$first_dose)
monoterapi$last_dose <- monoterapi$first_dose + months(24)
datahresept$datoAdministrasjonVirkestoff <- as.Date(datahresept$datoAdministrasjonVirkestoff, format = '%d%b%Y')
```

##Fjerner de som ikke har samme PID som vår gruppe

```r
#fjerner de med ikke matchende PID
datahresept1 <- datahresept %>%
  filter(PID %in% monoterapi$PID)
#Sjekker antall
hresept1 <- n_distinct(datahresept1$PID)
cat("Number of unique PIDs: ", hresept1, "\n")
```

##Legger inn start og sluttdato for pembro

```r
#Lager 1 rad med PID basert på første first_dose last_dose slik at det ikke blir mange dupliserte rader i hresept2. leftjoin funker ikke uten modifisering.
monoterapi_summary <- monoterapi %>%
  group_by(PID) %>%
  summarise(
    first_dose = min(first_dose, na.rm = TRUE),
    last_dose = min(last_dose, na.rm = TRUE)
  ) %>%
  ungroup()

# Now join this summary with datahresept1 to get first_dose
datahresept2 <- datahresept1 %>%
  left_join(monoterapi_summary[, c("PID", "first_dose")], by = "PID")
datahresept3 <- datahresept2 %>%
  left_join(monoterapi_summary[, c("PID", "last_dose")], by = "PID")

# Check the number of observations to ensure they haven't inflated
hresept3 <- n_distinct(datahresept3$PID)
cat("Number of unique PIDs: ", hresept3, "\n")
```

##Fjerner alt før oppstart

```r
# Filtrerer ut administeringer som er gitt over 1 måned før diagnosedatoen.
datahresept4 <- datahresept3 %>%
  filter(datoAdministrasjonVirkestoff >= (first_dose %m-% months(1)) | is.na(first_dose))
#sjekker antall pasienter
antallmatchende3<- length(unique(datahresept4$PID))
```

##markerer de som bruker annen behandling før sluttdato

Her har vi en del som har fått proteinkinasehemmere. Ser litt nærmere på det.

Fra metodevurderingen. Pembrolizumab som monoterapi er indisert til førstelinjebehandling av metastatisk ikke-småcellet lungekreft (NSCLC) hos voksne med tumor som uttrykker PD-

L1 med ≥50% «tumour proportion score» (TPS) uten EGFR- eller ALK-positive mutasjoner i tumor.

```
# Mark patients who receive other treatments before slutt_dato
datahresept5 <- datahresept4 %>%
  mutate(
    OtherBeforeEnd = ifelse(datoAdministrasjonVirkestoff < last_dose , TRUE,FALSE)
  )

# Now, you can filter changed treatment
changed_treatment <- datahresept5 %>%
  filter(OtherBeforeEnd)


# Check the results
head(changed_treatment)
```

##Teller de som er adherent/nonadherent

```
# Count the unique PIDs where OtherBeforeEnd is TRUE (non-adherence)
non_adherence_count <- datahresept5 %>%
  filter(OtherBeforeEnd == TRUE) %>%
  summarise(UniquePIDCount = n_distinct(PID))

# Count the unique PIDs where OtherBeforeEnd is FALSE (adherence)
adherence_count <- datahresept5 %>%
  filter(OtherBeforeEnd == FALSE) %>%
  summarise(UniquePIDCount = n_distinct(PID))

# Print the counts
print(non_adherence_count)
print(adherence_count)
```

##Deler de i 2 datasett

```
# Create two separate datasets based on OtherBeforeEnd value
non_adherence_dataset <- datahresept5 %>%
  filter(OtherBeforeEnd == TRUE)

adherence_dataset <- datahresept5 %>%
  filter(OtherBeforeEnd == FALSE)

# If you want to view or write these datasets to CSV files
# View the datasets
head(non_adherence_dataset)
head(adherence_dataset)
```

11.1 Non-adherent info

Veldig mye alektinib. Disse kan fjernes siden det ikke er samme type pasient som metodevurderingen bruker. Samme gjelder krizotinib. Noe av det har ikke indikasjon heller, kan være en annen type kreft som behandles også. Uansett må disse fjernes.

fra FK: alektinib: Monoterapi til førstelinjebehandling av voksne med anaplastisk lymfokinase (ALK)-positiv, avansert ikke-småcellet lungekreft (NSCLC). Monoterapi til behandling av voksne med ALK-positiv, avansert NSCLC tidligere behandlet med krizotinib.

osimertinib: til behandling av voksne med lokalavansert eller metastatisk EGFR T790M-mutasjonspositiv NSCLC.

metodevurderingen: tumor som uttrykker PD-L1 med ≥50% «tumour proportion score» (TPS) uten EGFR- eller ALK-positive mutasjoner i tumor.

```
virkestoff_summary <- non_adherence_dataset %>%
  group_by(virkestoff) %>%
  summarise(UniquePIDCount = n_distinct(PID)) %>%
  ungroup() %>%
  arrange(desc(UniquePIDCount))

# Print the summary table
print(virkestoff_summary)
```

11.2 adherent info

Mye vinorelbin, som er vanlig vedlikeholdsbehandling.Noen av PKA hemmerene har ikke indikasjon som vedlikeholdsbehandling, og enkelte ved mutasjon. Velger likevel å beholde disse siden de har stått på monoterapi i 2 år.

Et par PKA hemmere, men de er innenfor indikasjon ved vedlikeholdsbehandling: eksempel Tarceva: ikkeke-småcellet lungekreft (NSCLC): Førstelinjebehandling hos pasienter med lokalavansert eller metastatisk NSCLC med EGFR-aktiverende mutasjoner. Vedlikeholdsbehandling hos pasienter med lokalavansert eller metastatisk NSCLC med EGFR-aktiverende mutasjoner og stabil sykdom etter førstelinje kjemoterapi. Behandling av pasienter med lokalt fremskreden eller metastatisk NSCLC etter minst ett tidligere mislykket kjemoterapiregime. Hos pasienter med tumor uten EGFR-aktiverende mutasjoner er Tarceva indisert når andre behandlingsalternativer ikke anses som egnet.

```
#teller antall brukere per virkestoff
virkestoff_summary2 <- adherence_dataset %>%
  group_by(virkestoff) %>%
  summarise(UniquePIDCount = n_distinct(PID)) %>%
  ungroup() %>%
  arrange(desc(UniquePIDCount))

# Print the summary table
print(virkestoff_summary2)
```

##Fjerner nonadherents fra sykehusdata

```
before_exclusion_count <- length(unique(monoterapi$PID))
# Count unique PIDs in monoterapi before the anti-join
monoterapi_pid_count_before <- n_distinct(monoterapi$PID)
cat("Number of unique PIDs in monoterapi before removal: ", monoterapi_pid_count_before,
 "\n")

# Count unique PIDs in non_adherence_dataset
non_adherence_pid_count <- n_distinct(non_adherence_dataset$PID)
cat("Number of unique PIDs in non_adherence_dataset: ", non_adherence_pid_count, "\n")
```

```r
# Perform the anti-join to remove non-adherent PIDs from monoterapi
monoterapi_adherent <- anti_join(monoterapi, non_adherence_dataset, by = "PID")

# Count unique PIDs in monoterapi after the anti-join
monoterapi_pid_count_after <- n_distinct(monoterapi_adherent$PID)
cat("Number of unique PIDs in monoterapi after removal: ", monoterapi_pid_count_after, "\n")

after_exclusion_count <- length(unique(monoterapi_adherent$PID))

# Update log
exclusion_log <- rbind(exclusion_log, data.frame(
  Step = 5,
  Reason = "No aduvant treatment",
  ExcludedCount = before_exclusion_count - after_exclusion_count,
  RemainingCount = after_exclusion_count
))
```

##Fjerner nonadherents fra kreftdata

```r
# Count unique PIDs in inputertpdl1 before the removal


inputertpdl1_pid_count_before <- n_distinct(inputertpdl1$PID)
cat("Number of unique PIDs in inputertpdl1 before removal: ", inputertpdl1_pid_count_before, "\n")

# Perform the anti-join to remove non-adherent PIDs from inputertpdl1
realworldpasienter <- semi_join(inputertpdl1, monoterapi_adherent, by = "PID")

# Count unique PIDs in inputertpdl1 after the removal
inputertpdl1_pid_count_after <- n_distinct(realworldpasienter$PID)
cat("Number of unique PIDs in inputertpdl1 after removal: ", inputertpdl1_pid_count_after, "\n")
```

```r
# Now you can compare the before and after counts to ensure that the non-adherent PIDs have been removed

cat("Number of unique PIDs in monoterapi after removal: ", monoterapi_pid_count_after, "\n")
```

#Eksporterer filene

```r
# Skriver ut Hresept datasett
write.csv(non_adherence_dataset, "Hreseptnon_adherence.csv", row.names = FALSE)
write.csv(adherence_dataset, "Hreseptadherence_dataset.csv", row.names = FALSE)
write.csv(realworldpasienter, "realworldadherentmedavvik.csv", row.names = FALSE)
write.csv(monoterapi_adherent, "sykehusadherentmedavvik.csv", row.names = FALSE)
write.csv(exclusion_log, "exclusion_log2.csv", row.names = FALSE)
# Skriver ut sykehusdata med adherence

# Detect the computer's name
computer_name <- Sys.info()["nodename"]

# Set the working directory based on the computer's name
if (computer_name == "JEROEN-LAPTOP") {
  setwd("C:\\masteroppgaven lokal\\raw data\\studyvsrwd")
} else if (computer_name == "JEROENHAUKAAS") {
  setwd("D:/Masteroppgaven backup/raw data/studyvsrwd")
} else {
  stop("Unknown computer: unable to set the working directory")
}

# Load data
studypopulation <- read.csv("patientdatareck.csv", header = TRUE, sep = ",")
realworldpopulation <- read.csv("rwdpasientdatamedavvik.csv", header = TRUE, sep = ",")
weibull <- read.csv("weibull.csv", header = TRUE, sep = ",")
```

```r
library(tidyverse)
library(ggplot2)
library(janitor)
library(mice)
library(dplyr)
library(DataExplorer)
library(webshot2)
library(openxlsx)
library(tidyverse)
library(tidyr)
library(janitor)
library(rstatix)
library(remotes)
library(kableExtra)
library(devtools)
library(glmulti)
library(report)
library(sjPlot)
#library(ggstatsplot)
library(survival)
library(survminer)
library(biostat3)
library(tidyverse)
library(ggsurvfit)
library(dplyr)
library(gtsummary)
library(gridExtra)
library(scales)
library(survextrap)

## Warning: package 'survextrap' was built under R version 4.3.2
```

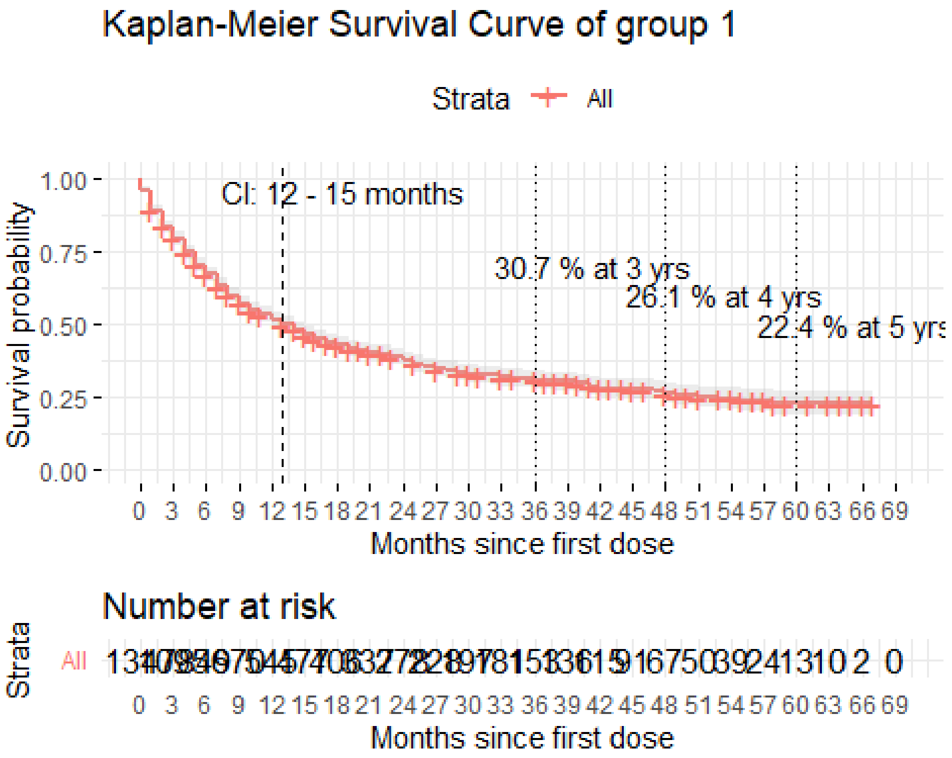
##Reck et al studie

```
km_fit <- survfit(Surv(time = studypopulation$V1, event = studypopulation$V2) ~ 1)

surv_object <- Surv(time = studypopulation$V1, event = studypopulation$V2)
survival_plot <- ggsurvplot(
  km_fit, data = studypopulation, conf.int = TRUE,
  risk.table = TRUE,
  xlab = "Months since first dose", ylab = "Survival probability",
  title = "Kaplan-Meier Survival Curve of group 1",
  ggtheme = theme_minimal(),
  break.x.by = 3,  # Set x-axis breaks every 10 months
  xlim = c(0, 70)  # Extend x-axis to 60 months
)
print(survival_plot)
```

##Vaar data

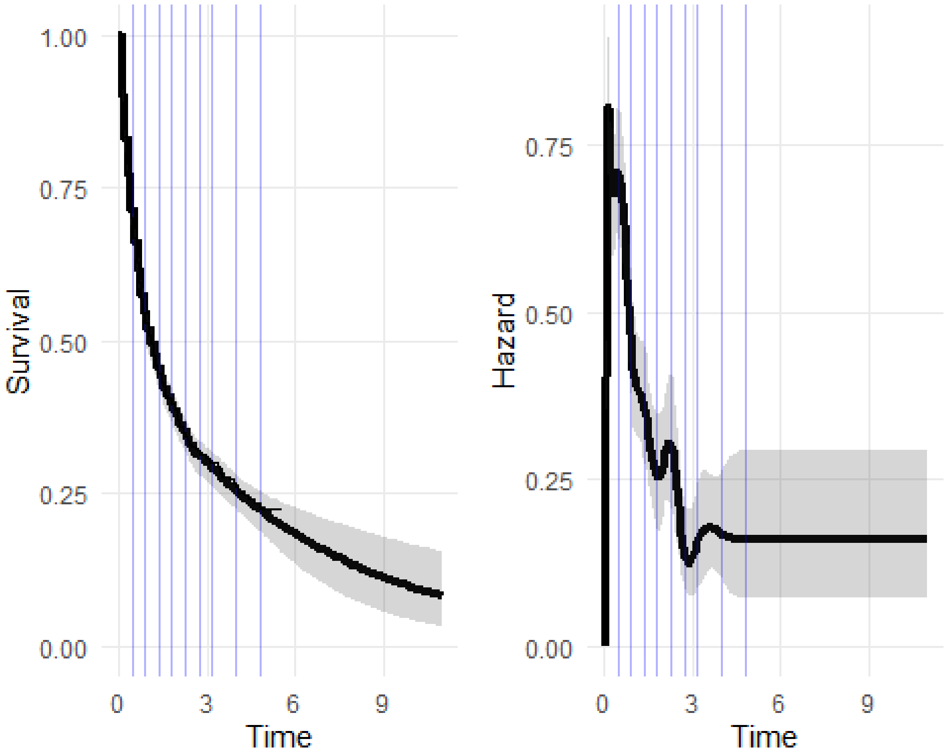
```
# Create the survival object using months
surv_obj <- Surv(time = realworldpopulation$survival_time_months, event = realworldpopulation$censoring_status)

# Fit the Kaplan-Meier survival curve
km_fit <- survfit(surv_obj ~ 1, data = realworldpopulation)

# Calculate survival probabilities at specific time points (36, 48, and 60 months)
time_points_months <- c(36, 48, 60)  # 3, 4, and 5 years in months
survival_probabilities <- summary(km_fit, times = time_points_months)$surv

# Extract survival probabilities
survival_at_3_years <- survival_probabilities[1]
survival_at_4_years <- survival_probabilities[2]
survival_at_5_years <- survival_probabilities[3]

# Find the median survival time in months
median_survival_months <- summary(km_fit)$table['median']
```

```r
# Extract the survival curve data
surv_data <- broom::tidy(km_fit, conf.int = TRUE)

# Calculate differences in months from the median to the lower and upper CI bounds
lower_ci_diff <- surv_data$time[which.max(surv_data$conf.low <= 0.5)]
upper_ci_diff <- surv_data$time[which.max(surv_data$conf.high <= 0.5)]

# Calculate the CI for the median survival time
lower_ci_median <- lower_ci_diff
upper_ci_median <- upper_ci_diff

survival_plot <- ggsurvplot(
  km_fit, data = realworldpopulation, conf.int = TRUE,
  risk.table = TRUE,
  xlab = "Months since first dose", ylab = "Survival probability",
  title = "Kaplan-Meier Survival Curve of group 1",
  ggtheme = theme_minimal(),
  break.x.by = 3,  # Set x-axis breaks every 10 months
  xlim = c(0, 70)  # Extend x-axis to 60 months
)

# Add median survival line and year markers
year_markers <- c(36, 48, 60)  # 3, 4, and 5 years in months
survival_plot$plot <- survival_plot$plot +
  geom_vline(xintercept = year_markers, linetype = "dotted", color = "black") +
  geom_vline(xintercept = median_survival_months, linetype = "dashed", color = "black")

# Offset for annotations
offset <- -5.5

# Annotate the median survival time and its CI
median_annotation <- paste("Median:", round(median_survival_months, 1), "months\nCI:",
                           round(lower_ci_median, 1), "-", round(upper_ci_median, 1), "months")
survival_plot$plot <- survival_plot$plot +
  annotate("text", x = median_survival_months - offset, y = 0.8, label = median_annotation, vjust = -0.5, color = "black") +
  annotate("text", x = 36 - offset, y = 0.7, label = paste(round(survival_at_3_years * 100, 1), "% at 3 yrs"), color = "black") +
  annotate("text", x = 48 - offset, y = 0.6, label = paste(round(survival_at_4_years * 100, 1), "% at 4 yrs"), color = "black") +
  annotate("text", x = 60 - offset, y = 0.5, label = paste(round(survival_at_5_years * 100, 1), "% at 5 yrs"), color = "black")

# Print and save the plot
print(survival_plot)
```



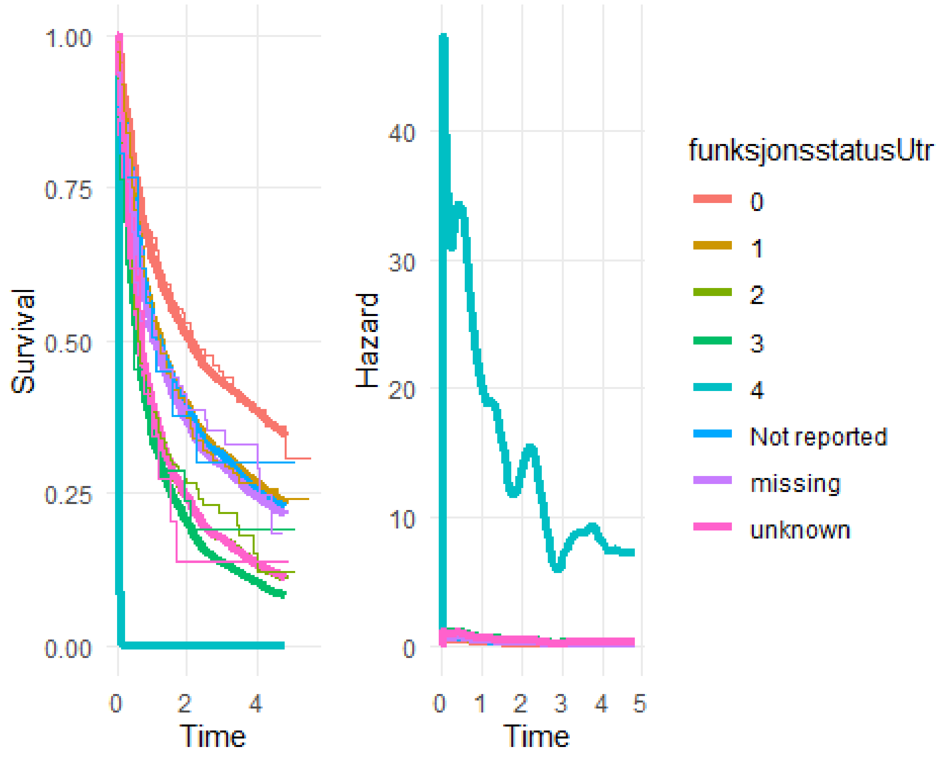
11.3 survextrap

```
# Copy the original dataframe to create a new dataset
realworldpopulation$survival_time_years <- realworldpopulation$survival_time_months / 12
new_realworldpopulation <- realworldpopulation

# Modify the survival_time_years in the new dataset
new_realworldpopulation$survival_time_years[new_realworldpopulation$survival_time_year
s == 0] <- 1/365



km_fit <- survfit(Surv(survival_time_years, censoring_status) ~ 1, data=new_realworldpopulation)

nd_mods <- survextrap(Surv(survival_time_years, censoring_status) ~ 1, chains=1, data=new_realworldpopulation)

##
## SAMPLING FOR MODEL 'survextrap' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 0.000923 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 9.23 seconds.
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 1:
## Chain 1:  Elapsed Time: 24.425 seconds (Warm-up)
## Chain 1:                21.249 seconds (Sampling)
## Chain 1:                45.674 seconds (Total)
## Chain 1:

plot(nd_mods,show_knots=TRUE, tmax=11)
```



```
rmst(nd_mods, t = c(10,15,20), niter=100)

## # A tibble: 3 × 5
##   variable      t median lower upper
##   <chr>     <dbl>  <dbl> <dbl> <dbl>
## 1 rmst         10   2.68  2.41  3.03
## 2 rmst         15   2.99  2.54  3.69
## 3 rmst         20   3.13  2.57  4.15

rxph_mod <- survextrap(Surv(survival_time_years, censoring_status) ~ funksjonsstatusUtr,
data=new_realworldpopulation, chains=1, refresh=0)

## Warning: Some Pareto k diagnostic values are too high. See help('pareto-k-diagnostic') for
details.
```

```
summary(rxph_mod) |>
  filter(variable=="loghr")

## # A tibble: 7 × 9
##   variable basis_num term    median   lower upper    sd  rhat ess_bulk
##   <chr>        <dbl> <chr>    <dbl>   <dbl> <dbl> <dbl> <dbl>    <dbl>
## 1 loghr           NA funksj…  0.352  0.150  0.559 0.102  1.00     523.
## 2 loghr           NA funksj…  0.779  0.557  1.01  0.113  1.01     480.
## 3 loghr           NA funksj…  0.891  0.524  1.22  0.176  1.00     671.
## 4 loghr           NA funksj…  4.17   0.0596 6.73  1.76   1.00     584.
## 5 loghr           NA funksj…  0.380 -0.240  0.888 0.304  1.00     729.
## 6 loghr           NA funksj…  0.408  0.0947 0.688 0.153  1.00     633.
## 7 loghr           NA funksj…  0.822  0.214  1.38  0.299  1.01     775.

plot(rxph_mod, niter=100)
```



```
nd <- data.frame(funksjonsstatusUtr = c("0","1","2","3","missing"))
rmst(rxph_mod, t=c(15), newdata=nd)
```

```
## # A tibble: 5 × 6
##   variable funksjonsstatusUtr     t median lower upper
##   <chr>    <chr>              <dbl>  <dbl> <dbl> <dbl>
## 1 rmst     0                     15   4.78 3.68   6.01
## 2 rmst     1                     15   3.20 2.59   4.11
## 3 rmst     2                     15   1.79 1.43   2.41
## 4 rmst     3                     15   1.49 0.945  2.48
## 5 rmst     missing               15   3.04 2.09   4.41
```

```
table(new_realworldpopulation$funksjonsstatusUtr)
```

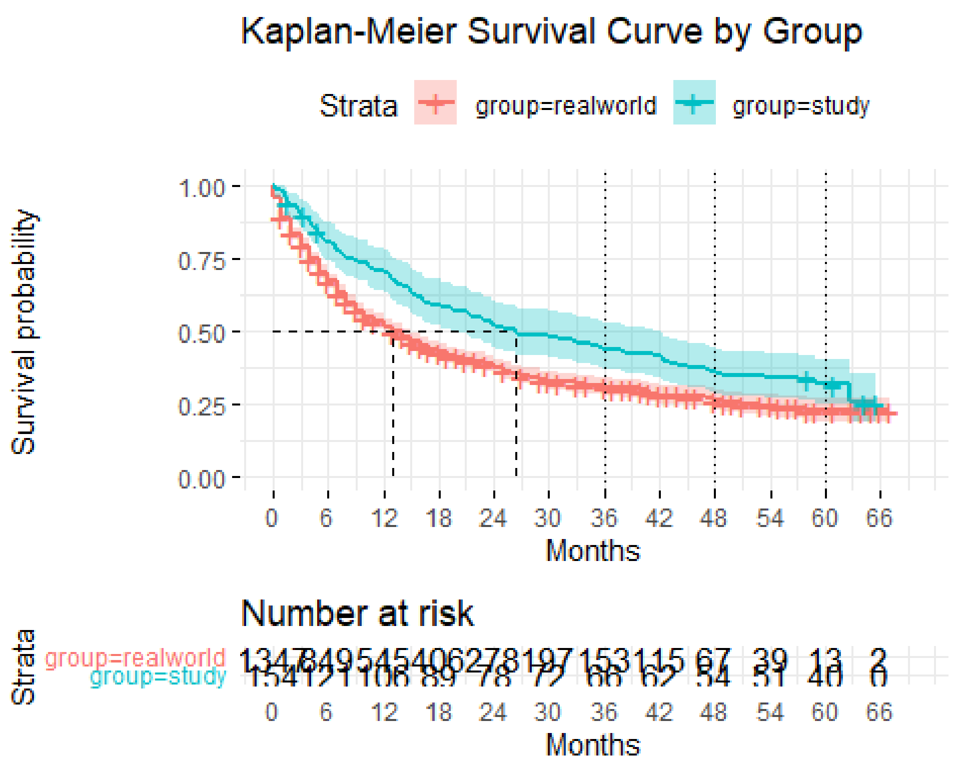
```
##
##            0            1            2            3            4
##          276          540          319           68            1
## Not reported      missing      unknown
##           26          101           16
```

##kombinerer til 1 datasett

```
names(studypopulation)[1] <- 'survival_time1'
names(studypopulation)[2] <- 'censoring_status'

# Then use the exact name for renaming
names(realworldpopulation)[names(realworldpopulation) == 'survival_time'] <- 'survival_timedays'
names(realworldpopulation)[names(realworldpopulation) == 'survival_time_months'] <- 'survival_time1'

studypopulation$V3 <- NULL
realworldpopulation <- realworldpopulation[, c('survival_time1', 'censoring_status')]

studypopulation$group <- 'study'
realworldpopulation$group <- 'realworld'

combined_data <- rbind(studypopulation, realworldpopulation)
```

```
# Assuming combined_data contains the variables 'survival_time', 'censoring_status', and 'group'

# Create the survival object
surv_obj_combined <- Surv(time = combined_data$survival_time, event = combined_data$censoring_status)

# Fit the Kaplan-Meier survival curve stratified by group
km_fit_combined <- survfit(surv_obj_combined ~ group, data = combined_data)

# Calculate survival probabilities at specific time points (36, 48, and 60 months)
time_points_months <- c(36, 48, 60)  # 3, 4, and 5 years in months
survival_probabilities <- summary(km_fit_combined, times = time_points_months)$surv
print(survival_probabilities)

## [1] 0.3066313 0.2612797 0.2244513 0.4381438 0.3584813 0.3161210

# Assuming the pattern is Group1 (36 months), Group2 (36 months), Group1 (48 months), Group2 (48 months), etc.
survival_at_3_years_group1 <- survival_probabilities[4]
survival_at_3_years_group2 <- survival_probabilities[1]
survival_at_4_years_group1 <- survival_probabilities[5]
survival_at_4_years_group2 <- survival_probabilities[2]
survival_at_5_years_group1 <- survival_probabilities[6]
survival_at_5_years_group2 <- survival_probabilities[3]



# Find the median survival time in months for each group
median_survival_months <- summary(km_fit_combined)$table['median']

# Extract the survival curve data
surv_data <- broom::tidy(km_fit_combined, conf.int = TRUE)

# Plotting the Kaplan-Meier curve
survival_plot_combined <- ggsurvplot(
  km_fit_combined, data = combined_data, conf.int = TRUE,
  risk.table = TRUE,
  xlab = "Months", ylab = "Survival probability",
  title = "Kaplan-Meier Survival Curve by Group",
  ggtheme = theme_minimal(),
  break.x.by = 6,  # Set x-axis breaks every 12 months
  xlim = c(0, 70),  # Extend x-axis to 70 months
  surv.median.line = "hv"
)

# Add median survival line and year markers
year_markers <- c(36, 48, 60)  # 3, 4, and 5 years in months
survival_plot_combined$plot <- survival_plot_combined$plot +
  geom_vline(xintercept = year_markers, linetype = "dotted", color = "black") +
  geom_vline(xintercept = median_survival_months, linetype = "dashed", color = "black")
annotations <- data.frame(
  time = c(36, 48, 60, 36, 48, 60)+3.5,
  survival = c(survival_at_3_years_group1, survival_at_4_years_group1, survival_at_5_years_group1,
               survival_at_3_years_group2, survival_at_4_years_group2, survival_at_5_years_group2),
  group = rep(c("Group 1", "Group 2"), each = 3),
  vjust = c(-6, -6, -6, 4, 6.5, 6.5)  # Example adjustments, modify as needed
)

# Add text annotations with adjusted vertical positions
survival_plot_combined$plot <- survival_plot_combined$plot +
  geom_text(data = annotations, aes(x = time, y = survival, label = paste(round(survival * 100, 1), "%"), vjust = vjust),
            color = "black",size = 6.3)

# Print and save the plot
print(survival_plot_combined)
```

```
## Warning: Removed 1 rows containing missing values (`geom_vline()`).
## Removed 1 rows containing missing values (`geom_vline()`).
```



```
# Combine the plot and risk table into a single grid object
km_combined_grid <- arrangeGrob(survival_plot_combined$plot, survival_plot_combined$table, ncol = 1, heights = c(5, 1))

## Warning: Removed 1 rows containing missing values (`geom_vline()`).

# Save the combined plot as an image
ggsave("studyvsrwd.png", km_combined_grid, width = 15, height = 11, bg = "#FFFDFB", dpi = 300)
```

##tester

```
# Create the survival object
surv_obj_combined <- Surv(time = combined_data$survival_time, event = combined_data$censoring_status)

# Perform the log-rank test
log_rank_test <- survdiff(surv_obj_combined ~ group, data = combined_data)
```

```r
# Output the results of the log-rank test
print(log_rank_test)

## Call:
## survdiff(formula = surv_obj_combined ~ group, data = combined_data)
##
##                    N Observed Expected (O-E)^2/E (O-E)^2/V
## group=realworld 1347      799      765      1.47      10.4
## group=study      154      107      141      8.01      10.4
##
##  Chisq= 10.4  on 1 degrees of freedom, p= 0.001

# Assuming you have the survfit object for each group from previous analysis (km_fit_combined)

# Calculate mean survival times for each group directly from the survfit summary
mean_survival_times <- summary(km_fit_combined)$table

# Print mean survival times
print(mean_survival_times)

##                 records n.max n.start events    rmean se(rmean)
## group=realworld    1347  1347    1347    799 25.61448 0.8346888
## group=study         154   154     154    107 33.63960 2.0988709
##                   median  0.95LCL  0.95UCL
## group=realworld 13.00000 12.00000 15.00000
## group=study     26.32988 19.40811 41.39491
```

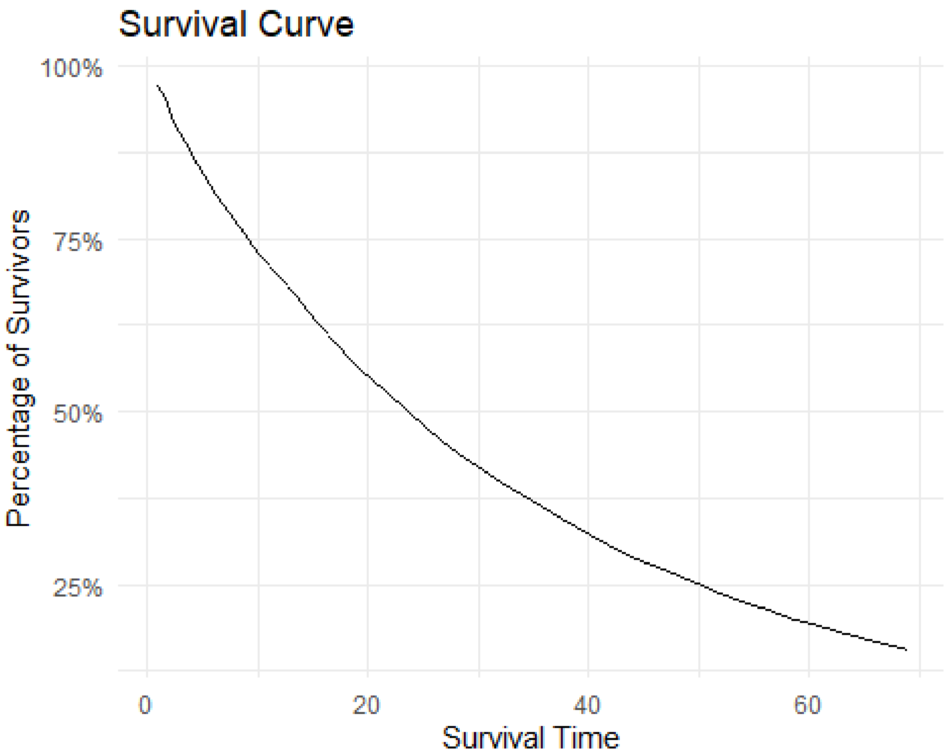
11.4 eksporterer snitt

```r
# Convert the matrix to a data frame
mean_survival_times_df <- as.data.frame(mean_survival_times)

# Add the row names as a new column in the data frame
mean_survival_times_df$group <- rownames(mean_survival_times_df)
```

```
# Reorder the columns to move the new 'group' column to the front, if desired
mean_survival_times_df <- mean_survival_times_df[, c('group', names(mean_survival_times_df)[1:9])]

# Load the openxlsx library
library(openxlsx)

# Write the data frame to an Excel file
write.xlsx(mean_survival_times_df, file = "mean_survival_times.xlsx")
```

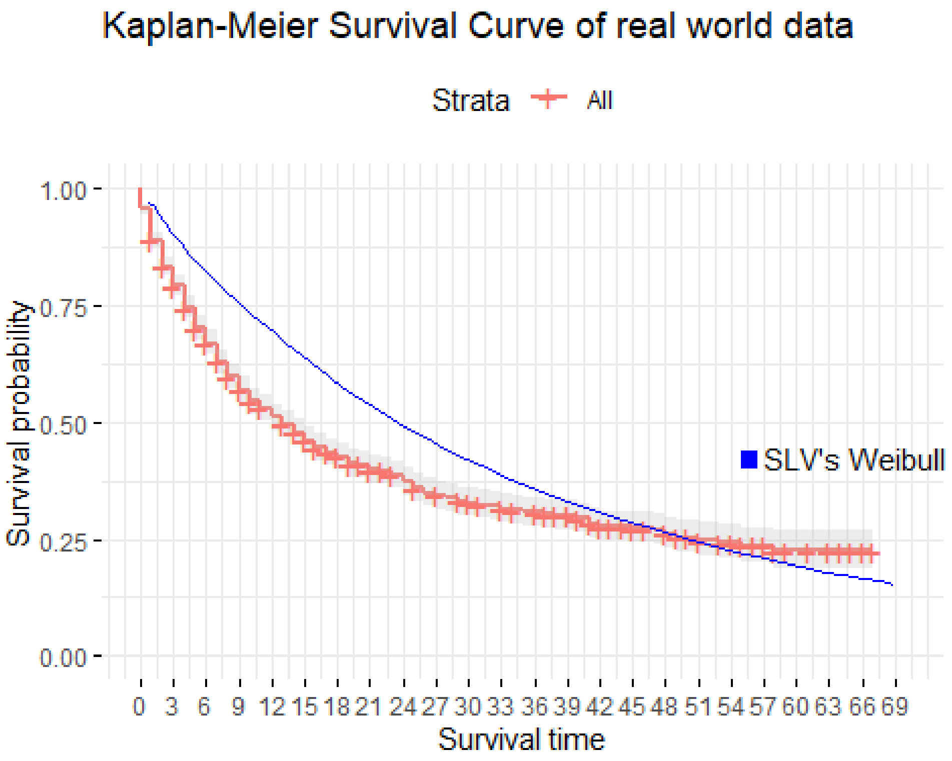
11.5 weibull fremskrivningen

```
names(weibull) <- c('Survival_time', 'Percentage_Survivors')
weibull$Survival_time <- weibull$Survival_time / 4.345
ggplot(weibull, aes(x = Survival_time, y = Percentage_Survivors)) +
  geom_line() +  # Draw the line
  scale_y_continuous(labels = scales::percent) +  # Convert y-axis to percentage
  labs(title = "Survival Curve", x = "Survival Time", y = "Percentage of Survivors") +
  theme_minimal()
```



##weibull + noinput

```r
# Fit the Kaplan-Meier survival curve
km_fit <- survfit(surv_obj ~ 1, data = realworldpopulation)

# Create the Kaplan-Meier plot
survival_plot <- ggsurvplot(
  km_fit, data = realworldpopulation, conf.int = TRUE,
  risk.table = FALSE,   # Do not show the risk table
  pval = FALSE,         # Do not show the p-value
  xlab = "Survival time", ylab = "Survival probability",
  title = "Kaplan-Meier Survival Curve of real world data",
  ggtheme = theme_minimal(),
  break.x.by = 3,      # Set x-axis breaks every 10 months
  xlim = c(0, 70),     # Extend x-axis to 60 months
  surv.misc = FALSE    # Remove miscellaneous information including 'Strata = all'
)

# Assuming your Weibull data is in a dataframe called weibull_data
# and the time is already converted to months (weibull_data$Survival_Time_Months)

# Overlay the Weibull function line on the Kaplan-Meier plot
survival_plot$plot <- survival_plot$plot +
  geom_line(data = weibull, aes(x = Survival_time, y = Percentage_Survivors), color = "blue") +
  geom_rect(aes(xmin = 55, xmax = 56.5, ymin = 0.4, ymax = 0.44), fill = "blue") +
  annotate("text", x = 57, y = 0.427, label = "SLV's Weibull model", color = "black",size=4, hjust = 0)



# Print and save the plot with the Weibull line
print(survival_plot)
```



```
ggsave("studywithweibull.png", survival_plot$plot, width = 15, height = 11, bg = "#FFFDFB
", dpi = 300)
```

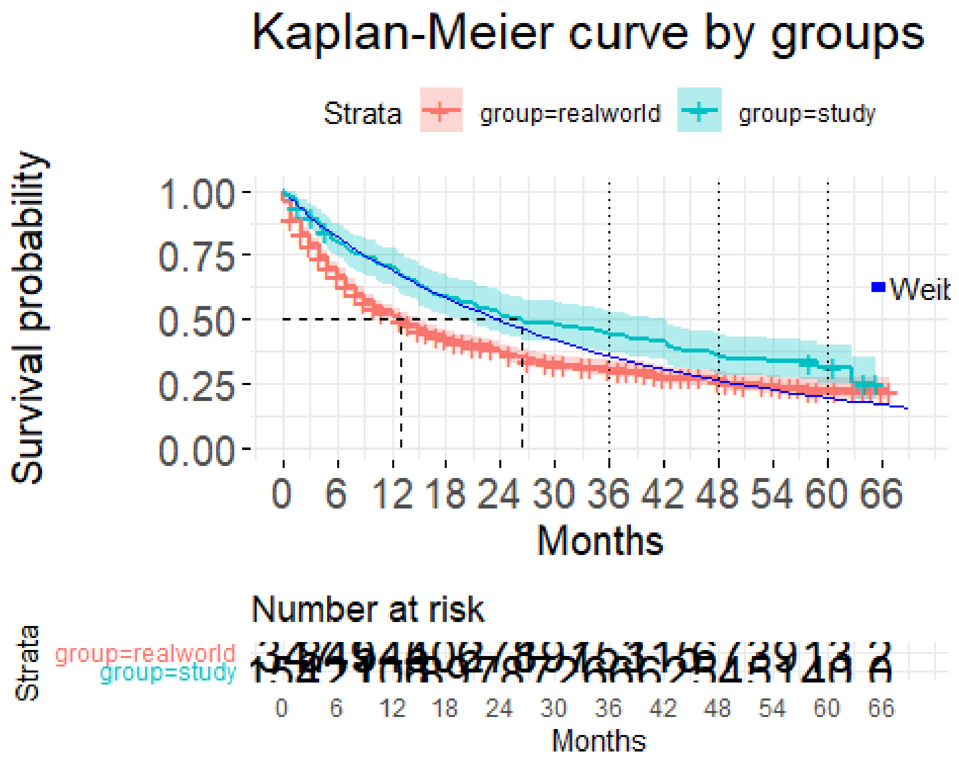
12 weibull, noinput og study

```
# Create the survival object
surv_obj_combined <- Surv(time = combined_data$survival_time, event = combined_data$c
ensoring_status)

# Fit the Kaplan-Meier survival curve stratified by group
km_fit_combined <- survfit(surv_obj_combined ~ group, data = combined_data)

# Calculate survival probabilities at specific time points (36, 48, and 60 months)
time_points_months <- c(36, 48, 60)  # 3, 4, and 5 years in months
survival_probabilities <- summary(km_fit_combined, times = time_points_months)$surv
print(survival_probabilities)
```

```
## [1] 0.3066313 0.2612797 0.2244513 0.4381438 0.3584813 0.3161210
```

```r
# Assuming the pattern is Group1 (36 months), Group2 (36 months), Group1 (48 months), Group2 (48 months), etc.
survival_at_3_years_group1 <- survival_probabilities[4]
survival_at_3_years_group2 <- survival_probabilities[1]
survival_at_4_years_group1 <- survival_probabilities[5]
survival_at_4_years_group2 <- survival_probabilities[2]
survival_at_5_years_group1 <- survival_probabilities[6]
survival_at_5_years_group2 <- survival_probabilities[3]



# Find the median survival time in months for each group
median_survival_months <- summary(km_fit_combined)$table['median']

# Extract the survival curve data
surv_data <- broom::tidy(km_fit_combined, conf.int = TRUE)

# Plotting the Kaplan-Meier curve
survival_plot_combined <- ggsurvplot(
  km_fit_combined, data = combined_data, conf.int = TRUE,
  risk.table = TRUE,
  xlab = "Months", ylab = "Survival probability",
  title = "Kaplan-Meier curve by groups",
  ggtheme = theme_minimal(),
  break.x.by = 6,  # Set x-axis breaks every 12 months
  xlim = c(0, 70),  # Extend x-axis to 70 months
  surv.median.line = "hv",
risk.table.fontsize = 6  # Adjust the font size of the numbers at risk (12 in this example)
)



# Add median survival line and year markers
year_markers <- c(36, 48, 60)  # 3, 4, and 5 years in months
survival_plot_combined$plot <- survival_plot_combined$plot +
  geom_vline(xintercept = year_markers, linetype = "dotted", color = "black") +
  geom_vline(xintercept = median_survival_months, linetype = "dashed", color = "black")
annotations <- data.frame(
  time = c(36, 48, 60, 36, 48, 60)+3.5,
  survival = c(survival_at_3_years_group1, survival_at_4_years_group1, survival_at_5_years_group1,
               survival_at_3_years_group2, survival_at_4_years_group2, survival_at_5_years_group2),
  group = rep(c("Group 1", "Group 2"), each = 3),
  vjust = c(-6, -6, -6, 4, 6.5, 6.5)  # Example adjustments, modify as needed
)

# Add text annotations with adjusted vertical positions
survival_plot_combined$plot <- survival_plot_combined$plot +
  geom_text(data = annotations, aes(x = time, y = survival, label = paste(round(survival * 100, 1), "%"), vjust = vjust),
            color = "black",size = 7.3)

# Add median survival line and year markers
year_markers <- c(36, 48, 60)  # 3, 4, and 5 years in months
survival_plot_combined$plot <- survival_plot_combined$plot +
  geom_line(data = weibull, aes(x = Survival_time, y = Percentage_Survivors), color = "blue") +
  geom_rect(aes(xmin = 65, xmax = 66.5, ymin = 0.6, ymax = 0.64), fill = "blue") +
  annotate("text", x = 67, y = 0.625, label = "Weibull model", color = "black", size = 4, hjust = 0)

survival_plot_combined$plot <- survival_plot_combined$plot +
  theme(
    plot.title = element_text(size = 19),  # Adjust the size of the title
    axis.title = element_text(size = 15),
    axis.text = element_text(size = 15))  # Adjust the size of the numbers at risk
```

```
# Print and save the plot
print(survival_plot_combined)

## Warning: Removed 1 rows containing missing values (`geom_vline()`).
## Removed 1 rows containing missing values (`geom_vline()`).
```



```
# Combine the plot and risk table into a single grid object
km_combined_grid <- arrangeGrob(survival_plot_combined$plot, survival_plot_combined$
table, ncol = 1, heights = c(5, 1))

## Warning: Removed 1 rows containing missing values (`geom_vline()`).

# Save the combined plot as an image
ggsave("studyvsrwdweibull.png", km_combined_grid, width = 15, height = 11, bg = "#FFFD
FB", dpi = 300)
```

##survextrap combo

```
# Copy the original dataframe to create a new dataset
newcombined <- combined_data
newcombined$survival_time1 <- newcombined$survival_time1 / 12
```

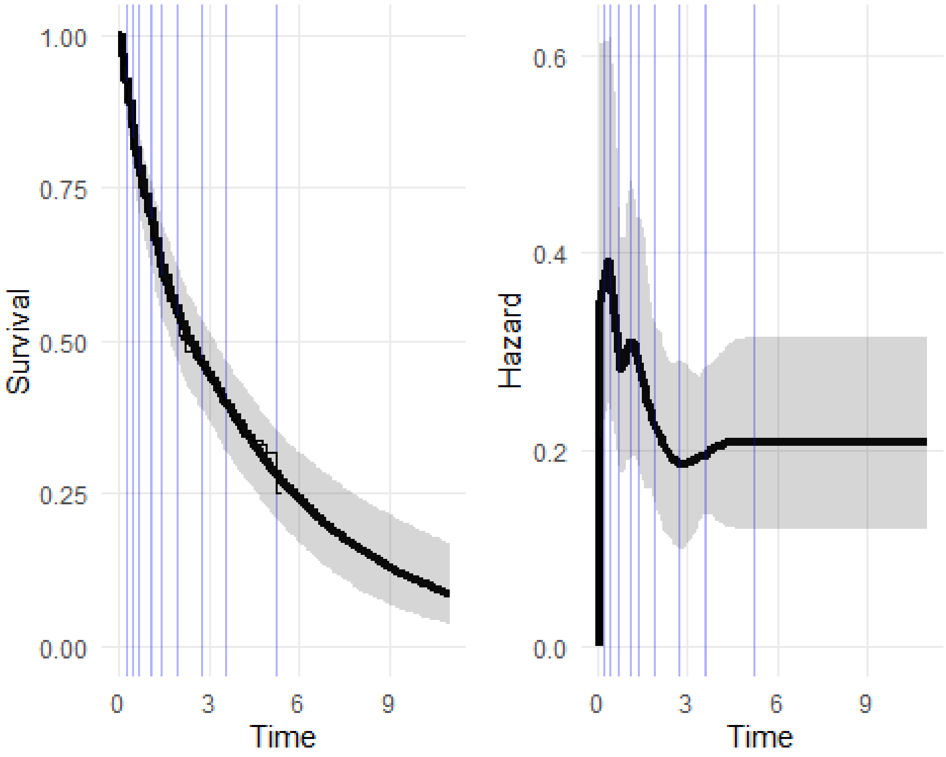
```
newcombined$survival_time1[newcombined$survival_time1 == 0] <- 1/365
newcombined$treat <- newcombined$group



# Fit separate models for each group if survextrap does not support stratification directly
nd_mod_study <- survextrap(Surv(survival_time1, censoring_status) ~ 1, data = newcombined[newcombined$treat == 'study', ], chains = 1)

## 
## SAMPLING FOR MODEL 'survextrap' NOW (CHAIN 1).
## Chain 1: 
## Chain 1: Gradient evaluation took 0.000177 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 1.77 seconds.
## Chain 1: Adjust your expectations accordingly!
## Chain 1: 
## Chain 1: 
## Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 1: 
## Chain 1:  Elapsed Time: 2.165 seconds (Warm-up)
## Chain 1:                1.562 seconds (Sampling)
## Chain 1:                3.727 seconds (Total)
## Chain 1: 
```

```
## Warning: There were 1 divergent transitions after warmup. See
## https://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
## to find out why this is a problem and how to eliminate them.

## Warning: Examine the pairs() plot to diagnose sampling problems
```

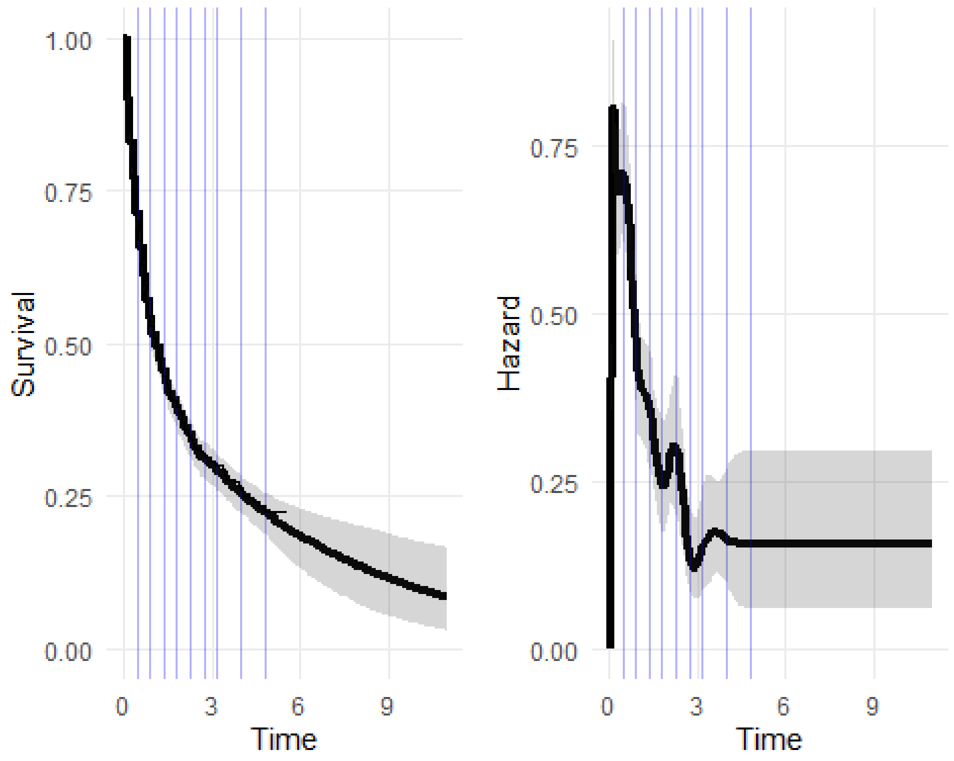
```
nd_mod_rwd <- survextrap(Surv(survival_time1, censoring_status) ~ 1, data = newcombined[newcombined$treat == 'realworld', ], chains = 1)
```

```
##
## SAMPLING FOR MODEL 'survextrap' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 0.000783 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 7.83 seconds.
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
## Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
## Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
## Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
## Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
## Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
## Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
## Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
## Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
## Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
## Chain 1:
## Chain 1:  Elapsed Time: 24.862 seconds (Warm-up)
## Chain 1:                19.404 seconds (Sampling)
## Chain 1:                44.266 seconds (Total)
## Chain 1:
```

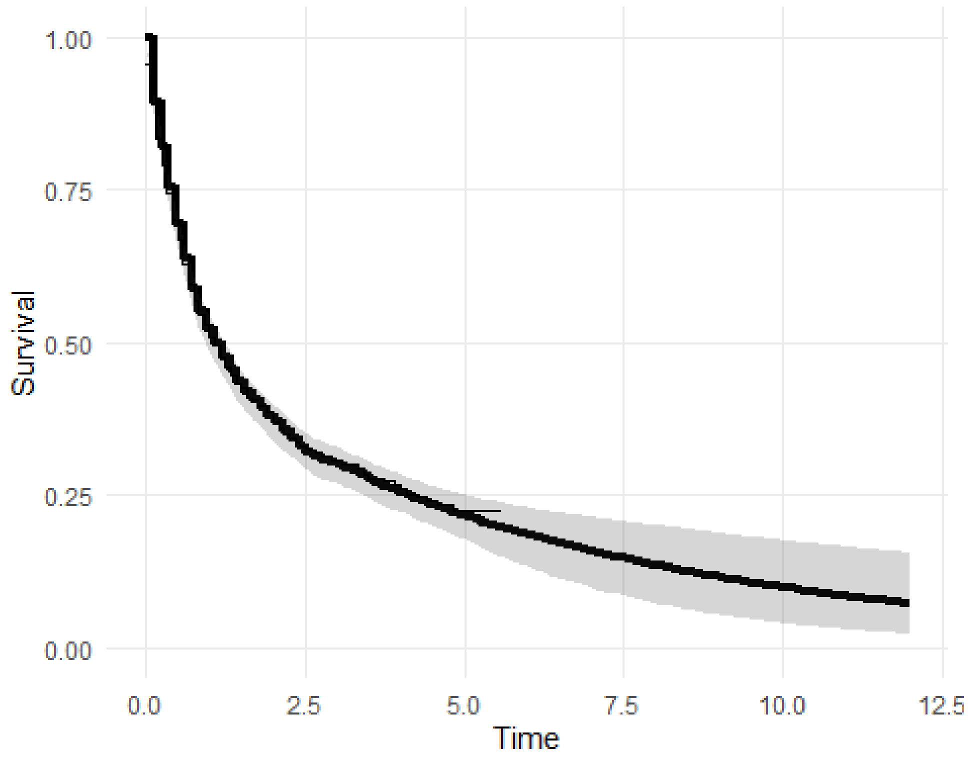
```
## Warning: There were 1 divergent transitions after warmup. See
## https://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
## to find out why this is a problem and how to eliminate them.

## Warning: Examine the pairs() plot to diagnose sampling problems

# Plot the extrapolations for each group
plot(nd_mod_study, show_knots = TRUE, tmax = 11)
```



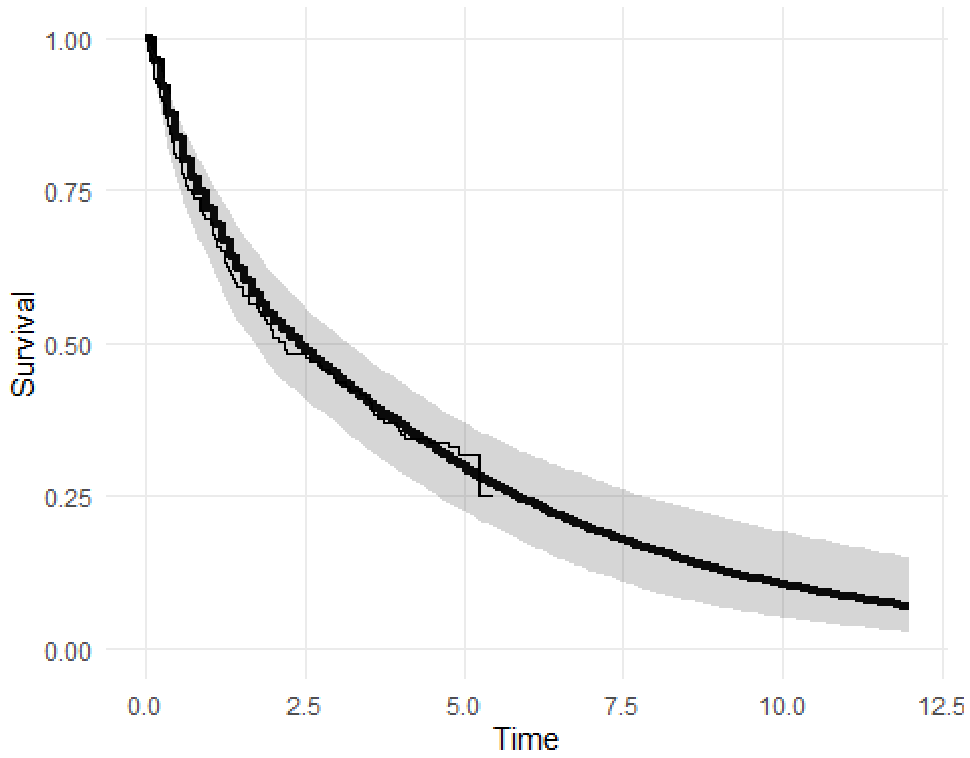
```
plot(nd_mod_rwd, show_knots = TRUE, tmax = 11)
```



```
plot_survival(nd_mod_rwd,tmax=12)
```



```
plot_survival(nd_mod_study,tmax=12)
```



~~~
```r
# Detect the computer's name
computer_name <- Sys.info()["nodename"]

# Set the working directory based on the computer's name
if (computer_name == "JEROEN-LAPTOP") {
  setwd("C:\\masteroppgaven lokal\\raw data\\studyvsrwd")
} else if (computer_name == "JEROENHAUKAAS") {
~~~

```
  setwd("D:/Masteroppgaven backup/raw data/studyvsrwd")
} else {
  stop("Unknown computer: unable to set the working directory")
}

# Load data
studypopulation <- read.csv("patientdatareck.csv", header = TRUE, sep = ",")
realworldpopulation <- read.csv("rwdpasientdatamedavvik.csv", header = TRUE, sep = ",")
sykehus <- read.csv("Utlevert_mkb_sykehus_4082.csv", header = TRUE, sep = ";")

library(tidyverse)
library(ggplot2)
library(janitor)
library(mice)
library(dplyr)
library(DataExplorer)
library(webshot2)
library(openxlsx)
library(tidyverse)
library(tidyr)
library(janitor)
library(rstatix)
library(remotes)
library(kableExtra)
library(devtools)
library(glmulti)
library(report)
library(sjPlot)
#library(ggstatsplot)
library(survival)
library(survminer)
library(biostat3)
library(tidyverse)
library(ggsurvfit)
library(dplyr)
library(gtsummary)
library(gridExtra)
library(scales)
```

##filtrerer ut

```
sykehus_filtered <- sykehus %>%
  filter(PID %in% realworldpopulation$PID)
pembrolizumab_data <- sykehus_filtered %>%
  filter(virkestoff == "Pembrolizumab")

# Counting unique PIDs
unique_pids_count <- sykehus_filtered %>%
                    summarise(count = n_distinct(PID)) %>%
                    pull(count)

# Display the count
unique_pids_count

# Calculate average dosage for Pembrolizumab
pembrolizumab_data_adjusted <- pembrolizumab_data %>%
  mutate(doseVirkestoff = ifelse(doseVirkestoff == 0, 200, doseVirkestoff))

totaldosage <- pembrolizumab_data_adjusted %>%
  summarise(TotalDose = sum(doseVirkestoff, na.rm = TRUE))

pembrolizumab_data_adjusted <- pembrolizumab_data %>%
  mutate(doseVirkestoff = ifelse(doseVirkestoff > 399, doseVirkestoff / 2, doseVirkestoff))
averagedosage <- pembrolizumab_data_adjusted %>%
  summarise(AverageDose = mean(doseVirkestoff, na.rm = TRUE))

library(openxlsx)
```

```
# Combine the data frames
combined_dosage <- bind_rows(totaldosage, averagedosage)
write.xlsx(combined_dosage, file = "Combined_Dosage_Pembrolizumab.xlsx")
```

child=‘D:/masteroppgaven backup/raw data/0 inputtering/Med avvik.Rmd’}

child=‘D:/masteroppgaven backup/raw data/0 inputtering/datarenssykehus.Rmd’}

child=‘D:/masteroppgaven backup/raw data/0 inputtering/datarenshresept med avvik.Rmd’}

child=‘D:/masteroppgaven backup/raw data/studyvsrwd/noinputvsstudy.Rmd’}

child=‘D:/masteroppgaven backup/raw data/studyvsrwd/doser gitt.Rmd’}

UiT Norges arktiske universitet