MASTER THESIS

A comparison of the quality of breast cancer care in Norway and the Netherlands

DAVE T. HAMERSMA
S2081326

FACULTY OF SCIENCE AND TECHNOLOGY
prof. dr. J.L. HEREK

EXAMINATION COMMITTEE
prof. dr. S. Siesling
dr. C.G.M. Groothuis-Oudshoorn

09-12-2020

Abstract

Introduction

Breast cancer is the most common cancer and one of the leading causes of death among women. To support the delivery of the highest quality of care provided by hospitals in Europe to women with breast cancer, the European Society of Breast Cancer Specialists defined quality indicators that act as a quality instrument for hospitals to standardize the quality assurance of these hospitals and set a standard minimum of care. Comparing quality indicators amongst countries may identify areas for improvement, opens discussions and further improve the quality of breast cancer care. In this study, comparisons were made of two geographically different countries.

Methods

Anonymized data was gathered from the Netherlands Cancer Registry and the Cancer Registry of Norway. The data selected was grouped in two populations, all female invasive breast cancer patients diagnosed in 2017 and 2018 in the Netherlands and all female invasive breast cancer patients diagnosed in 2017 and 2018 in Norway. Five European Society of Breast Cancer Specialists quality indicators were selected for assessment. Two based on MRI availability, two on appropriate surgical approaches and one on post-operative radiotherapy. The quality indicator outcomes were calculated before and after a federated Propensity Score Stratification on the two populations to reduce the bias of confounding by indication.

Results

In total 39,163 female breast cancer patients were included. 32,786 from the Netherlands and 6377 from Norway. The balance did improve after Propensity Score Stratification of every quality indicator. The outcome of the first MRI availability quality indicator were in the Netherlands 37% and Norway 17.5%. The second MRI availability was in the Netherlands 83.3% and Norway 70.8%. The first quality indicator of the appropriate surgical approach was in the Netherlands 95.2% and Norway 91.5%. The second in the Netherlands 36% and Norway 37.4%. Lastly, the quality indicator on post-operative radiotherapy was in the Netherlands 94.9% and Norway 95.7%.

Conclusion

In both countries four of five quality indicators were well above the minimum standard set by EUSOMA. The main differences between the countries are attributed to the implementation time of the guidelines. Both countries offer a high quality of breast cancer care compared to other countries and may yet improve even more in the future.

Keywords: breast cancer care, quality indicators, quality of care

Acknowledgements

Throughout the data gathering, writing of code, and writing of this thesis I have received a great deal of assistance and support.

Firstly, I would like to thank my supervisor, prof. dr. Sabine Siesling, who was there to offer help when needed and assist me in reaching out to contributors. I would also like to thank dr. Karin Groothuis-Oudshoorn as well, for helping me out on the data science challenges.

I wish to thank the data-managers and employees from Norway Cancer Registry as well, especially the head of Registry Informatics, Jan Nygård, who made sure the gathering of data from Norway went as smoothly as possible. Due to the corona crisis, I was unable to spend a few months in Norway to meet everyone and work on this thesis. I will visit Norway in another time and hope to experience the life there.

I would like to thank the data-managers and employees of Cancer Registry Netherlands, and in particular Erik Heeg and Kay Schreuder, for answering all the questions I had, their insights and their attention (even late in the evening). I would also like to thank the data science team of Cancer Registry Netherlands for supporting and including me in their team. Gijs Geleijnse for his support and attention from the beginning to the end, Frank Martin for his exceptional computer science skills and Matteo Cellamare for his incredible mathematical skills.

Enjoy reading,

Dave T. Hamersma

Introduction

Breast cancer is the most common cancer and one of the leading causes of death among women (1). To support the delivery of the highest quality of care provided by hospitals in Europe to women with breast cancer, the European Society of Breast Cancer Specialists (EUSOMA) was founded in 1986. EUSOMA defined quality indicators that act as guidelines for hospitals to standardize the quality assurance of these hospitals and set a standard minimum of care (2). These quality indicators aim to cover every aspect of the cancer care process, from diagnosis to surgery and treatment. In total EUSOMA defined thirty-four benchmark quality indicators with seventeen categories. These categories include the assessment of, diagnosis, surgery, treatment, and rehabilitation. Hospitals can participate voluntary in the EUSOMA to apply for a Breast Centre Certification by submitting data and discuss the indicators during an audit visit, which is possible to apply for every two years (3). When a hospital wants to receive the status "Specialist Breast Cancer", the hospital needs to achieve the minimum standard of fourteen out of the seventeen categories of quality indicators set by EUSOMA (2). Furthermore, this EUSOMA standard enables hospitals to compare their own hospital with other hospitals within the individual country. Comparing and evaluating hospitals' quality indicators between hospitals and countries are an advised method to further improve the quality of care (4).

However, comparisons of countries are challenging since the differences might be influenced by other underlying characteristics and sharing sensitive patient data might pose difficulties. The data owned by European Union countries are affected by the General Data Protection Regulation (GDPR), which introduces restrictions on data sharing due to potential privacy sensitive data leaks (5). However, the Netherlands Comprehensive Cancer Registration Organisation (IKNL) has developed an open-source federated learning infrastructure, Personal Health Train (PHT), where sites using the infrastructure share their statistical model and model parameters instead of sharing sensitive data (6). By incorporating PHT, comparisons can be made in coherence with GDPR and thus, without sharing sensitive patient data.

In this study, comparisons were made of two geographical different countries, Norway and the Netherlands. Norway is considered one the most sparsely populated countries in Europe (7), while the Netherlands is one of the most densely populated country in Europe (8). This means that accessibility to hospitals differ greatly between these countries. All Dutch people live within twenty-five minutes of a hospital (8). In Norway there are more individual differences in access to hospitals, with in the most rural part, hospitals are located with 500 kilometers of one another (7). However, most Norwegian hospitals are in urban areas. The current population of the Netherlands is 17.4 million (9), Norway's population is 5.4 million (10). Despite the differences of the countries, both strive for a good quality of care. In relation to the differences in breast cancer, the incidence of breast cancer diagnoses in the Netherlands was in 2019 14,808 invasive breast cancer and 2,229 in-situ breast cancer (11), of all cancer cases 28% were breast cancer amongst women (12). In Norway in 2018 of all new cancer cases, 22.3% or 3,568 women were diagnosed with breast cancer (13). The five-year relative survival of breast cancer stage combined in Norway was 90.7% in 2018 (13), while the Netherlands the average five-year survival rate is 87% (11). Both Norway and the Netherlands have similar biennial mammography breast cancer screening programs. However in the Netherlands women are screened between the ages 50 and 74 (14), while Norway's screening program are between the ages 50 and 70 (13).

The differences in incidence, patient characteristics and geography could be indicating that there are different strategies and levels of expertise in the breast cancer care process within the individual country. With the fact that both countries strive for a high quality of care, the aim of this study is to gain insight in the differences of the quality of breast cancer care in the countries and enabling the ability to learn from each other by evaluating EUSOMA's quality indicators.

Methods

Anonymized data was gathered from the Netherlands Cancer Registry (NCR) and the Cancer Registry of Norway (CRN). Both cancer registries are covering the complete population. The NCR is hosted by the IKNL, which has data managers in all hospitals collecting data directly from the patient files based on a notification by the Automated Pathology Archive (PALGA) (15). CRN collects data of all cancer cases, which is based on reports by medical doctors in Norway (16). These reports are sent at different times; at the time of the diagnosis, each surgical event, primary adjuvant treatment, the start of hormone therapy and the end of hormone therapy (17).

The data was collected and grouped in two populations, all female invasive breast cancer patients from the Netherlands diagnosed in 2017 and 2018 and all female invasive breast cancer patients from Norway diagnosed in 2017 and 2018.
For the assessment of quality of care within countries, quality indicators defined by EUSOMA were selected for comparison. Due to availability of data, relevancy, and clinical importance the EUSOMA quality indicators presented in table 1 were selected for assessment.

Table 1: the selected EUSOMA indicators

EUSOMA quality indicators (2)	
MRI availability: 6a	
Numerator	Number of patients that was examined preoperatively by magnestic resonance imaging (MRI)
Denominator	Number of patients that received an operation
Exclusion	Patients with PST
Minimum standard	10%
MRI availability: 6b	
Numerator	Number of patients treated with PST undergoing MRI (pre, during, post PST)
Denominator	Number of patients treated with PST
Exclusion	Patients with distant metastasis
Minimum standard	60%
Appropriate surgical approach: 9a	
Numerator	Number of patients who received a single breast operation for primary tumour
Denominator	Number of patients that received an operation
Exclusion	Patients that underwent a reconstruction
Minimum standard	80%
Appropriate surgical approach: 9c	
Numerator	Number of patients that received an immediate reconstruction at the same time of mastectomy
Denominator	Number of patients that received a mastectomy

Exclusion	None
Minimum standard	40%
Post-operative radiotherapy: 10a	
Numerator	Number of patients who received postoperative radiation therapy after surgical resection of the primary tumour and appropriate axillary staging/surgery in the framework of breast conserving therapy
Denominator	Number of patients with surgical resection of the primary tumour and appropriate axillary staging/surgery in the framework of breast conserving therapy
Exclusion	Patients with distant metastasis
Minimum standard	90%

To adjust for differences in patient characteristics, Propensity Score Stratification (PSS) was used to balance the two countries. PSS is a technique used in observational studies to reduce bias from confounding by indication, by stratifying the data in *k* number of strata based on a 'propensity score'. This propensity score is calculated with a generalized linear regression and a log link function with the country as the dependent variable and the independent variables the potential confounders. The interpretation of a propensity score would be that the probability of assignment to a country based on the baseline characteristics of that patient. When using PSS, a large portion of the original sample size will be retained (18) and with at least 5 strata, 90% of the bias can be removed (19). The PSS was applied on each quality indicator and within a federated learning infrastructure (Personal Health Train), with both countries' dataset located at the respective owner. In the appendix is a full description of the PSS in a federated infrastructure supplemented.

One of the challenges of a propensity score calculation between countries, is that in the potential confounders (independent variables) there could be differences in ways of registration or in definition. In table 2 the definitions of patient characteristics that were provided in the data exchange and used as independent variables in the calculation of the propensity score are clarified.

Table 2: Definitions of independent variables

Independent variable	**The Netherlands (15)**	**Norway (20)**
Year of diagnosis	Year of the incidence date, first date when the tumor/relapse/progression was diagnosed	The first date where the diagnosis is confirmed
Age	Age of patient at the year of diagnosis	Age of patient at the year of diagnosis
Histological tumor type	Derived from the ICD-O-3 morphology code	Derived from the ICD-O-3 morphology code
Differentiation grade	Description of abnormality of tumor cells	Description of abnormality of tumor cells
Pathological T-stage (pT)	Pathological T-stage based on UICC TNM. Received before the (neoadjuvant) therapy, supplemented with information from (post-	Pathological T-stage based on UICC TNM. Derived from the pathology report.

	surgery) pathology examination	
Pathological N-stage (pN)	Pathological N-stage based on UICC TNM. Received before the (neoadjuvant) therapy, supplemented with information from (post-surgery) pathology examination	Pathological N-stage based on UICC TNM. Derived from the pathology report.
HER2 status	Her2 status measured by immunohistochemistry: -0-1+: Negative -3+: Positive -2+: Unknown	Her2 status measured by immunohistochemistry: -0-1+: Negative -3+: Positive -2+: Unknown
Estrogen receptor status	Estrogen receptor level before chemotherapy: -0-9%: Negative -10+%: Positive	Estrogen receptor level in tumor: -<1%: Negative ->1%: Positive
Progesterone receptor status	Progesterone receptor level before chemotherapy: -0-9%: Negative -10+%: Positive	Progesterone receptor level in tumor: -0-9%: Negative -10+%: Positive

The balance of the data was calculated before PSS and after PSS with a Standardized Mean Difference (SMD) on every independent variable of each quality indicator. The SMD is one of the most commonly used statistics in propensity score studies to assess balance, with a higher value of 0.1 or lower value of -0.1 indicating imbalance (21). It is applicable to all variables due to the independency of unit of measurement (21). Since PSS divides the data in *k*-strata the SMD is applied across each stratum. If the balance did not improve for the specified independent variables, the number of strata is adjusted to finer or rougher strata. However, if any of the independent variables were known to be unrelated to the quality indicator, they were omitted to reduce noise. When greater balance is achieved, a quality indicator analysis was then performed. The quality indicator analysis was computed as an Average Treatment Effect, this means that the quality indicator will be calculated within each stratum defined by the PSS. Afterwards, the average will be calculated with a 95% confidence interval to achieve less biased quality indicator results. Finally, an odds ratio (OR) will be calculated across strata to define the differences in results.

Results

The data of the Netherlands consists of 32,786 female invasive breast cancer patients diagnosed in hospitals between 2017 to 2018 registered by the NCR. The CRN included 6377 female invasive breast cancer patients diagnosed between 2017 and 2018. The mean age for the Netherlands was 62.4 (SD ± 13.8) and for Norway 60.9 (SD ± 12.9). The descriptive analysis of the total populations is presented in table 3. The descriptive analysis of the subpopulations (every quality indicator) is given in Appendixes A through E. Before the analysis, the independent variable "differentiation grade" a level ("undifferentiated") and its population was completely removed due to low occurrence (n = 5) and the fact that it is not used clinically. Due to differences in registration, the level "no evidence of primary tumour" of independent variable "pT" was transformed to "Unknown" for the Netherlands.

Table 3: Descriptive analysis

	Norway (N=6377)	The Netherlands (N=32786)
Year of Diagnosis		
2017	3230 (50.7%)	16567 (50.5%)
2018	3147 (49.3%)	16219 (49.5%)
Age		
<40	342 (5.4%)	1758 (5.4%)
40-49	938 (14.7%)	4479 (13.7%)
50-59	1630 (25.6%)	7614 (23.2%)
60-69	1807 (28.3%)	8329 (25.4%)
70-79	1152 (18.1%)	6653 (20.3%)
80+	508 (8.0%)	3953 (12.1%)
Histological tumor type		
Ductal	4975 (78.0%)	25146 (76.7%)
Lobular	791 (12.4%)	4292 (13.1%)
Other	611 (9.6%)	3348 (10.2%)
Differentiation grade		
Well differentiated	1372 (21.5%)	7156 (21.8%)
Moderately differentiated	2789 (43.7%)	15434 (47.1%)
Poorly differentiated	1515 (23.8%)	7336 (22.4%)
Unknown	701 (11.0%)	2860 (8.7%)
pT		
Tumor size <2cm	3711 (58.2%)	18430 (56.2%)
Tumor size 2-5cm	1573 (24.7%)	6751 (20.6%)
Tumor size 5+ cm	104 (1.6%)	1142 (3.5%)
Unknown	989 (15.5%)	6463 (19.7%)
pN		
No regional lymph node metastasis	3941 (61.8%)	19520 (59.5%)
Metastasis in 1-3 lymph nodes	1508 (23.6%)	6684 (20.4%)
Metastasis in 4+ lymph nodes	237 (3.7%)	1261 (3.8%)
Unknown	691 (10.8%)	5321 (16.2%)
HER2 status		
Negative	5464 (85.7%)	27376 (83.5%)
Positive	829 (13.0%)	4168 (12.7%)
Unknown	84 (1.3%)	1242 (3.8%)
Estrogen receptor status		
Negative	906 (14.2%)	5011 (15.3%)
Positive	5393 (84.6%)	27417 (83.6%)
Unknown	78 (1.2%)	358 (1.1%)
Progesterone receptor status		
Negative	1944 (30.5%)	10100 (30.8%)
Positive	4358 (68.3%)	22306 (68.0%)
Unknown	75 (1.2%)	380 (1.2%)

MRI availability 1: pre-operative MRI

For the analysis of the quality indicator, 21,664 patients from the Netherlands and 5,262 patients from Norway were included. The full descriptive analysis table is provided in Appendix A. Before the analysis, variable pT was slightly adjusted, the level "Unknown" was removed due to low occurrence and interference with PSS. The level consists in the Netherlands of 161 patients (0.7%) and in Norway of 32 patients (0.6%). Before PSS, age, differentiation grade, pN and HER2 status had a higher SMD than the threshold of -0.1/0.1, which indicates a state of imbalance of the two countries. After applying a five strata PSS, the SMD's of these five imbalanced variables were significantly reduced and moved below the threshold. The quality indicator results in the Netherlands were 36.9% before stratification and 37% (95% CI 34.1-40) after (graph 1). In Norway, before stratification it was 18% and 17.5% (95% CI 15.3-19.7) after. The OR to be examined preoperatively by MRI in the Netherlands is 2.8 (95% CI 2.7-2.9) compared to Norway.

MRI availability 2: MRI during PST

The analysis of the quality indicator consists of 7,003 patients from the Netherlands and 752 from Norway. The full descriptive analysis table is provided in Appendix B. Variable pT and pN were removed and not incorporated in the PSS, due to differences in registration. Age, histological tumor type, differentiation grade, ER receptor status and PR receptor status had an SMD higher than the threshold. A five strata PSS resulted in a representable balance. With only year of diagnosis being over the threshold. However, the strata were not perfectly distributed with patients in Norway, with only 29 (4%) patients in stratum 5. Nonetheless, this did not influence the average results of the quality indicator. The quality indicator results of Norway were before stratification 75.3% and after 70.8% (95% CI 66.4-75.2) (graph 1). the Netherlands had before 83.8% and after stratification 83.3% (95% CI 79.1-87.5). The OR to undergo MRI with PST in the Netherlands is 2.3 (95% CI 1.3-3.3) compared to Norway.

Appropriate surgical approach 1: single breast operation

The first quality indicator of appropriate surgical approach included 28,806 patients from the Netherlands and 5,029 patients from Norway. The full descriptive analysis table is provided in Appendix C. Differentiation grade, pT and pN were imbalanced before the PSS. After applying a five strata PSS, only one pT was still imbalanced with an SMD of 0.101. Adjusting the number of strata did not further improve balance. The quality indicator results for Norway were before stratification 92% and after 91.5% (95% CI 89.1-93.9) (graph 1). Results from the Netherlands were before 95.2% and after stratification 95.2% (95% CI 94.5-95.9). The OR to receive a single breast operation in the Netherlands is 1.8 (95% CI 1.4-2.2) compared to Norway.

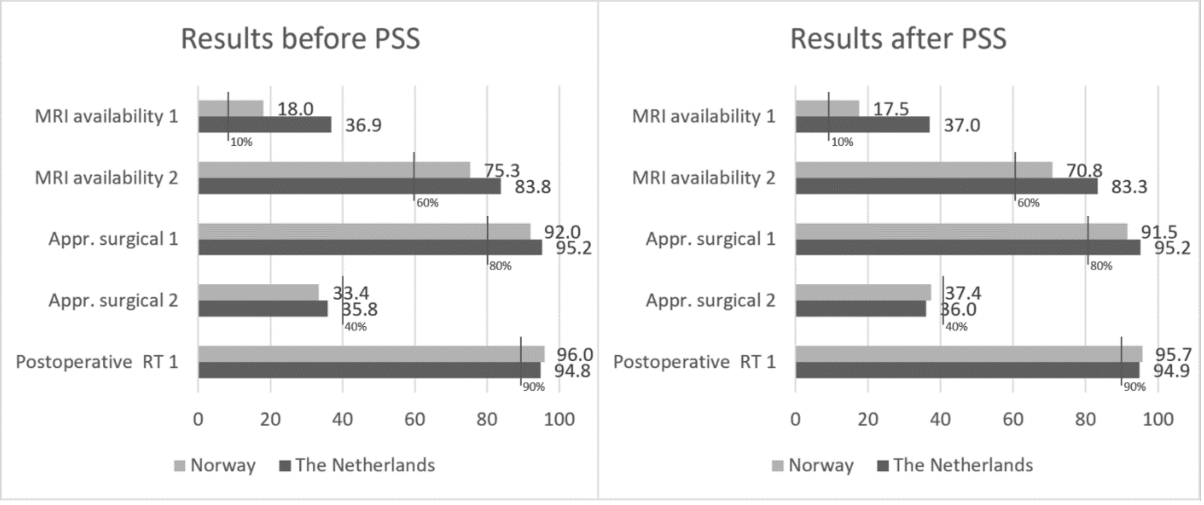
Appropriate surgical approach 2: immediate reconstruction

In this quality indicator 7,116 patients from the Netherlands and 748 from Norway were included. The full descriptive analysis table is provided in Appendix D. Differentiation grade, pT, pN and PR receptor status were imbalanced with an SMD higher than the threshold. The five strata PSS did not improve the balance of the data. The PSS was adjusted into finer strata, which improved the balance significantly. After a seven strata PSS, only differentiation grade had an SMD of 0.381. The results for QI 9c were before stratification for Norway 33.4% and after 37.4% (95% CI 29.8-44.9) (graph 1). For the Netherlands before 35.8% and after stratification 36% (95% CI 31.3-40.7). The OR to receive immediate reconstruction at the same time of mastectomy in the Netherlands is 1.2 (95% CI 0.7-1.7) in compared to Norway.

Post-operative radiotherapy 1: after surgical resection

In the analysis of the quality indicator 17,594 patients from the Netherlands and 3,748 patients from Norway were included in the analysis. The full descriptive analysis table is provided in Appendix E. Differentiation grade and pT were imbalanced before the PSS. This QI required a nine strata PSS to achieve a good balance and resulted that none of the variables had a SMD higher than the threshold. The results for this QI were for Norway before stratification 96% and after 95.7% (95% CI 94.6-96.7) (graph 1). For the Netherlands, the outcome was before stratification 94.8% and after 94.9% (95% CI 91.8-98). The OR to receive postoperative radiation therapy in the Netherlands is 1.1 (95% CI 0.8-1.5) compared to Norway.

Graph 1: Results EUSOMA Quality Indicators before and after PSS



Discussion

The aim of this study was to gain insight in the differences in the breast cancer care between the Netherlands and Norway so that it would enable the ability to learn from the results. As presented in this study, four out of five quality indicators were well over the minimum standard set by EUSOMA. Only the second quality indicator of Appropriate surgical approach was slightly below the minimum standard for both countries. After reducing the bias from confounding by indication, there were significant differences between the results of EUSOMA's quality indicators between countries. Notably in the MRI availability category, the first quality indicator is the Netherlands almost 20% (19.5%) higher than Norway, with an OR of 2.8. The first quality indicator of MRI availability relates to the percentage of patients that were examined preoperatively by MRI. However, due to the fact that this QI excludes patients with PST the clinical importance is reduced, and it acts more as a descriptive QI for information about the risk of overdiagnosis (2). In both the Norwegian and Dutch guidelines, the use of MRI is only recommended for selected patient groups (22, 23). However, these selected patient groups vary from each other as they are based on different literature. The significant difference in results could also be explained by the time of implementation in the breast cancer guidelines. In 2011, the Netherlands introduced new indications for preoperative MRI's in the breast cancer guideline (24). It states that patients with lobular invasive breast cancer are indicated to receive a preoperative MRI (22), as this reduces the percentage of reoperation and mastectomy (25, 26). The same indication was introduced in the Norwegian guidelines in 2017 (27). Since the data used in this study is from 2017 and 2018, it could be that the new guidelines were not fully adapted yet in Norway. It is noteworthy that there was an increase in QI results in Norway from 2017 to 2018, 16.7% to 19.3% respectively. It can be concluded that the results are mainly due to differences in clinical practices and may improve over time.

With the second quality indicator of MRI availability, which includes only patients treated with PST, the QI results differ 12.5% in favour of the Netherlands with an OR of 2.3. These results may be influenced from registration artefacts, since it became apparent that there are differences in ways of registration between Norway and the Netherlands. After a patient receives neoadjuvant primary systemic therapy in Norway, the pathology TNM classifications are not registered in the pathology report but as a new variable, which was not included in this study. This caused problems with the analysis and therefore, the pathology TNM classifications were removed from the analysis. Due to this obstacle, stratifying on the propensity score was less comprehensive. However, the difference is significant and could be explained by other factors. The motivation for undergoing MRI with PST, as defined by EUSOMA, is to proper evaluate the response to PST (2). In the Netherlands, this viewpoint has been introduced in the breast cancer guidelines since 2011 (28). Norway has introduced this since 2007 (29). Nonetheless, the percentage of patients undergoing an MRI with PST have been steadily increasing in the recent years in the Netherlands (28) and in Norway (30).

In the category appropriate surgical approach, both countries' QI results are similar. The first quality indicator differs 3.7% in favour of the Netherlands with an OR of 1.8, and the second differs 1.4% in favour of Norway with an OR of 1.2. With the first QI, both countries achieve the target determined by EUSOMA and have a considerable low reoperation rate compared to other European countries. When comparing the Norwegian reoperation rate of 8.4% to other Scandinavian countries, it is higher than Denmark (17%) (31) and Iceland (13.6-14.1%) (32), and similar to Finland (8.4%) (33). The reoperation rate of the Netherlands is significantly lower than other European countries, as it has been for multiple years (28). This can be attributed to the early implementation of this indication in the guidelines.

The second QI, which relates to the percentage of patients receiving an immediate reconstruction at the same time of mastectomy, is for both countries under the EUSOMA standard of 40%. However, compared to other countries Norway and the Netherlands are significantly superior (34-36). The Netherlands have more than doubled the percentage of patients receiving an immediate reconstruction at the same time of mastectomy, in 2011-2014 this was 17% (28) and now, presented in this study, 36%. The advice to perform a direct reconstruction at the same time of mastectomy has been indicated since the first breast cancer guideline of the Netherlands in 2002 (37). The first breast cancer guideline of Norway introduced in 2007 the notion that the cosmetic results may be just as good or better with an immediate reconstruction after mastectomy (29). It was in 2013 that the advice was added in the guideline to offer every female patient that undergo a mastectomy an immediate reconstruction (38). In 2016 the percentage of Norway was 27% (39) and now, as presented in this study, it is 37.4%. It seems that Norway is adapting the indication in the guideline slightly faster than the Netherlands. In the recent years, more and more studies have proven that immediate reconstruction after mastectomy provides positive effects, such as cosmetic satisfaction (40) and an increase of quality of life (41). Nonetheless, it is apparent in that in both countries younger patients are more likely to apt for immediate reconstruction than older patients. There are also other patient specific factors contributing to the QI results, the patient may not desire an immediate reconstruction or is unable to due to contraindications. Both countries' results are moving in the right direction, it could be that the percentage may be already over the minimum standard set by EUSOMA at this moment.

Norway and the Netherlands both achieved high results in the percentage of patients receiving post-operative radiotherapy, with only 0.8% difference between the countries and well over the minimum standard set by EUSOMA. The breast cancer guidelines of both countries present similar indications for patients to receive post-operative radiotherapy (22, 23). However, this quality indicator may never be fully 100%, as there are contraindications for the post-operative radiotherapy and in the end, the patient decision to receive the treatment. The results of the two countries are higher than the minimum standard (90%), but the percentage of Norway may even be higher than presented. The reason could be due to loss of registration since a hospital is offering an intraoperative radiation therapy. This experimental partial radiation therapy, which is delivered during the surgery, is usually indicated for patients with small tumours or patients that are unable to undergo the traditional postoperative therapy (42). This type of therapy is by the definition of EUSOMA, not considered post-operative radiotherapy but should, in fact, be included in the calculation. In the complete definition, provided by EUSOMA, is stated that "appropriate" axillary staging/surgery should be offered. In this case "appropriate" could be interpreted in various ways but, after consultation with EUSOMA, "appropriate" means that the patients are characterized by a known lymph node staging. This is noteworthy, since there was no specific information provided with the calculation of the quality indicators. The exact definition is still open for interpretation. In this study, the definitions were repeatedly checked amongst clinicians of both countries to present clear comparable results.

With the PSS, it was possible to increase the balance in each subpopulation of the quality indicator. In every subpopulation the differentiation grade and TNM classification variables were unbalanced, based on the SMD's. The PSS reduced the SMD's of most of the variables. However, the quality indicator results did differ only slightly. The second QI of appropriate surgical approach and MRI availability in Norway was corrected the most, with an increase and decrease of almost 4%. The differences in results after PSS in the Dutch subpopulations were low, with percentages of 0.5%. The effects of PSS on the data used in this study did alter the results of the quality indicators slightly for Norway.

Unfortunately, only five out of the thirty-six EUSOMA quality indicators could be calculated. Data gathered were not sufficient to calculate the other thirty-one quality indicators. Due to the way the data was gathered and structured, there were some limitations in the calculations of the quality indicators. For instance, the interpretations of the quality indicators itself were somewhat divided, as was apparent in the second MRI availability QI and the Post-operative radiotherapy QI. However, with good communication between countries the interpretation should be the same and results can be compared. Some of the variables itself were divided as well, as was the case with the pathology reports. The ER variable is slightly different in Norway than the Netherlands as well. In Norway if a patient has an estrogen receptor level of more than 1%, it is defined as "positive", in the Netherlands it is positive if the estrogen receptor level is 10% or above. This could have influenced the calculation of the propensity score and the distribution of the strata. The balance did improve after PSS of every QI, but the QI results before and after were similar. This could have been due to the fact that the PSS has been deployed in its most straightforward way; it could have been improved by methods of trimming or weighing (43).

For further studies, additional EUSOMA quality indicators and data of recent years, should provide a more comprehensive view of the quality of breast cancer care. And additionally, could identify more areas for improvement, open discussions further and improving the quality of care for breast cancer patients. In the two countries four of five EUSOMA quality indicators were well above the minimum standard. The main differences in the results are attributed to the implementation time of the guidelines. As presented in this study, both countries offer a high quality of breast cancer care.

References

1. Momenimovahed Z, Salehiniya H. Epidemiological characteristics of and risk factors for breast cancer in the world. Breast Cancer: Targets and Therapy. 2019;11:151.

2. Biganzoli L, Marotti L, Hart CD, Cataliotti L, Cutuli B, Kühn T, et al. Quality indicators in breast cancer care: An update from the EUSOMA working group. European Journal of Cancer. 2017;86:59-81.

3. Biganzoli L, Cardoso F, Beishon M, Cameron D, Cataliotti L, Coles CE, et al. The requirements of a specialist breast centre. The Breast. 2020.

4. Organization WH. Improving healthcare quality in Europe: Characteristics, effectiveness and implementation of different strategies: World Health Organization. Regional Office for Europe; 2019.

5. van Veen E-B. Observational health research in Europe: understanding the General Data Protection Regulation and underlying debate. European Journal of Cancer. 2018;104:70-80.

6. Moncada Torres A, Martin F, Sieswerda M, Soest J, Geleijnse G. VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange2020.

7. Ringard Å, Sagan A, Saunes I, Lindahl A. Norway: health system review. 2013.

8. Kroneman M, Boerma W, van den Berg M, Groenewegen P, de Jong J, van Ginneken E. Netherlands: health system review. 2016.

9. CBS. Population counter: CBS; 2020 [17/12/2020]. Available from: https://www.cbs.nl/en-gb/visualisations/population-counter.

10. SSB. Population: Statistik sentralbyra; 2020 [17/12/2020]. Available from: https://www.ssb.no/en/befolkning/statistikker/folkemengde/aar-per-1-januar.

11. RIVM. Breast cancer in the Netherlands: National Institute for Public Health and the Environment; [14/12/2020]. Available from: https://www.rivm.nl/en/breast-cancer-screening-programme/breast-cancer-in-netherlands.

12. Registry NC. Cijfers over kanker: IKNL; [14/12/2020]. Available from: www.cijfersoverkanker.nl.

13. IK Larsen BM, TB Johannesen, TE Robsahm, TK Grimsrud, S Larønningen, E Jakobsen, G Ursin. Cancer in Norway 2018-Cancer incidence, mortality, survival and prevalence in Norway. Cancer Registry of Norway; 2019.

14. Kelly de Ligt, Marianne Luyendijk, Marissa van Maaren, Linda de Munck, Kay Schreuder, Sabine Siesling, et al. borstkanker in Nederland trends 1989-2017. IKNL; 2018.

15. IKNL. Netherlands Cancer Registry (NCR): IKNL; [14/12/2020]. Available from: https://www.iknl.nl/en/ncr.

16. Norway CRo. About the Cancer Registry: kreftregisteret; [14/12/2020]. Available from: https://www.kreftregisteret.no/en/General/About-the-Cancer-Registry/.

17. Hartmann-Johnsen OJ, Kåresen R, Schlichting E, Naume B, Nygård JF. Using clinical cancer registry data for estimation of quality indicators: Results from the Norwegian breast cancer registry. International journal of medical informatics. 2019;125:102-9.

18. Guo S, Fraser MW. Propensity score analysis: Statistical methods and applications: SAGE publications; 2014.

19. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics. 1968:295-313.

20. Kreftregisteret. ELVIS 2018 [cited 2020 14/12/2020]. Available from: https://metadata.kreftregisteret.no/.

21. Zhang Z, Kim HJ, Lonjon G, Zhu Y. Balance diagnostics after propensity score matching. Annals of translational medicine. 2019;7(1).

22. NABON V. Richtlijn Mammacarcinoom. Landelijke richtlijn. Oncoline; 2020.

23. NBCG N. Nasjonalt handlingsprogram med retningslinjer for diagnostikk, behandling og oppfølging av pasienter med brystkreft: The Norwegian Directorate of Health,; 2017. 2020.

24. Nederland NBO. Richtlijn mammacarcinoom 2008. Vereniging van Integrale Kankercentra, Amsterdam. 2008.

25. Mann RM, Hoogeveen YL, Blickman JG, Boetes C. MRI compared to conventional diagnostic work-up in the detection and evaluation of invasive lobular carcinoma of the breast: a review of existing literature. Breast cancer research and treatment. 2008;107(1):1-14.

26. Lobbes MB, Vriens IJ, van Bommel AC, Nieuwenhuijzen GA, Smidt ML, Boersma LJ, et al. Breast MRI increases the number of mastectomies for ductal cancers, but decreases them for lobular cancers. Breast cancer research and treatment. 2017;162(2):353-64.

27. Kreftregisteret. Årsrapport 2017 med resultater og forbedringstiltak fra Nasjonalt kvalitetsregister for brystkreft. . Oslo: Kreftregisteret; 2018.

28. van Bommel AC, Spronk PE, Vrancken Peeters MJT, Jager A, Lobbes M, Maduro JH, et al. Clinical auditing as an instrument for quality improvement in breast cancer care in the Netherlands: The national NABON Breast Cancer Audit. Journal of surgical oncology. 2017;115(3):243-9.

29. Helsedirektoratet. Nasjonalt handlingsprogram med retningslinjer for diagnostikk, behandling og oppfølging av pasienter med brystkreft. 2007. Contract No.: IS-1524.

30. Kreftregisteret. Årsrapport 2019 med resultater og forbedringstiltak fra Nasjonalt kvalitetsregister for brystkreft. . Oslo: Kreftregisteret; 2020.

31. Bodilsen A, Bjerre K, Offersen BV, Vahl P, Ejlertsen B, Overgaard J, et al. The influence of repeat surgery and residual disease on recurrence after breast-conserving surgery: a Danish Breast Cancer Cooperative Group Study. Annals of surgical oncology. 2015;22(3):476-85.

32. Palsdottir E, Lund S, Asgeirsson K. Oncoplastic breast-conserving surgery in Iceland: a population-based study. Scandinavian Journal of Surgery. 2018;107(3):224-9.

33. Niinikoski L, Leidenius MH, Vaara P, Voynov A, Heikkilä P, Mattson J, et al. Resection margins and local recurrences in breast cancer: comparison between conventional and oncoplastic breast conserving surgery. European Journal of Surgical Oncology. 2019;45(6):976-82.

34. Liederbach E, Sisco M, Wang C, Pesce C, Sharpe S, Winchester DJ, et al. Wait times for breast surgical operations, 2003–2011: a report from the National Cancer Data Base. Annals of surgical oncology. 2015;22(3):899-907.

35. Lang JE, Summers DE, Cui H, Carey JN, Viscusi RK, Hurst CA, et al. Trends in post-mastectomy reconstruction: A SEER database analysis. Journal of surgical oncology. 2013;108(3):163-8.

36. Zhong T, Fernandes KA, Saskin R, Sutradhar R, Platt J, Beber BA, et al. Barriers to immediate breast reconstruction in the Canadian universal health care system. Journal of clinical oncology. 2014;32(20):2133-41.

37. Rutgers E, Nortier J, Tuut M, Van Tienhoven G, Struikmans H, Bontenbal M, et al. CBO-richtlijn'behandeling van het mammacarcinoom'. Nederlands Tijdschrift voor Geneeskunde. 2002;146(45):2144-51.

38. Helsedirektoratet. Nasjonalt handlingsprogram med retningslinjer for diagnostikk, behandling og oppfølging av pasienter med brystkreft. 2013. Contract No.: IS-2063.

39. Ripsrud IM. Subkutan mastektomi med primær rekonstruksjon og strålebehandling 2017.

40. Jagsi R, Li Y, Morrow M, Janz N, Alderman A, Graff J, et al. Patient-reported quality of life and satisfaction with cosmetic outcomes after breast conservation and mastectomy with and without reconstruction: results of a survey of breast cancer survivors. Annals of surgery. 2015;261(6):1198.

41. Teo I, Reece GP, Christie IC, Guindani M, Markey MK, Heinberg LJ, et al. Body image and quality of life of breast cancer patients: influence of timing and stage of breast reconstruction. Psycho-Oncology. 2016;25(9):1106-12.

42. Kalakota K, Small Jr W. Intraoperative radiation therapy techniques and options for breast cancer. Expert Review of Medical Devices. 2014;11(3):265-73.

43. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. PloS one. 2011;6(3):e18174.

Appendices

Appendix A: Results Quality Indicator 6a

	Yes		No		Before PSS	After PSS
Indicator 6a	**Norway**	**The Netherlands**	**Norway**	**The Netherlands**	**SMD**	**SMD**
	(N=947)	**(N=7995)**	**(N=4315)**	**(N=13669)**		
Year of Diagnosis						
2017	444 (46.9%)	4053 (50.7%)	2212 (51.3%)	7125 (52.1%)	0.022	0.002
2018	503 (53.1%)	3942 (49.3%)	2103 (48.7%)	6544 (47.9%)	-0.022	-0.002
Age						
<40	77 (8.1%)	445 (5.6%)	142 (3.3%)	184 (1.3%)	-0.068	-0.014
40-49	224 (23.7%)	1392 (17.4%)	434 (10.1%)	936 (6.8%)	-0.055	0.005
50-59	294 (31.0%)	2252 (28.2%)	1094 (25.4%)	3000 (21.9%)	-0.049	0.014
60-69	242 (25.6%)	2148 (26.9%)	1358 (31.5%)	4236 (31.0%)	-0.020	0.015
70-79	104 (11.0%)	1538 (19.2%)	869 (20.1%)	3783 (27.7%)	**0.148**	-0.028
80+	6 (0.6%)	220 (2.8%)	418 (9.7%)	1530 (11.2%)	0.001	-0.003
Histological tumor type						
Ductal	636 (67.2%)	5079 (63.5%)	3485 (80.8%)	11553 (84.5%)	-0.037	0.003
Lobular	252 (26.6%)	2111 (26.4%)	350 (8.1%)	788 (5.8%)	0.059	-0.003
Other	59 (6.2%)	805 (10.1%)	480 (11.1%)	1328 (9.7%)	-0.013	0.000
Differentiation grade						
Well differentiated	196 (20.7%)	1965 (24.6%)	1074 (24.9%)	4132 (30.2%)	0.091	0.009
Moderately differentiated	506 (53.4%)	4536 (56.7%)	2040 (47.3%)	6277 (45.9%)	0.031	-0.004
Poorly differentiated	228 (24.1%)	1339 (16.7%)	1137 (26.3%)	3004 (22.0%)	**-0.140**	-0.008
Unknown	17 (1.8%)	155 (1.9%)	64 (1.5%)	256 (1.9%)	0.028	0.011
pT						
1	610 (64.4%)	5408 (67.6%)	3030 (70.2%)	10232 (74.9%)	0.066	0.010
2	310 (32.7%)	2245 (28.1%)	1221 (28.3%)	3107 (22.7%)	-0.099	-0.014
3	27 (2.9%)	342 (4.3%)	64 (1.5%)	330 (2.4%)	0.089	0.012
pN						
0	669 (70.6%)	5494 (68.7%)	3186 (73.8%)	9845 (72.0%)	-0.055	-0.026
1	223 (23.5%)	1952 (24.4%)	859 (19.9%)	2616 (19.1%)	0.013	0.004
2+	46 (4.9%)	286 (3.6%)	173 (4.0%)	364 (2.7%)	-0.063	-0.018
Unknown	9 (1.0%)	263 (3.3%)	97 (2.2%)	844 (6.2%)	**0.168**	0.071
HER2 status						
Negative	817 (86.3%)	7279 (91.0%)	3801 (88.1%)	12086 (88.4%)	0.051	-0.031
Positive	114 (12.0%)	557 (7.0%)	469 (10.9%)	1193 (8.7%)	**-0.102**	-0.008
Unknown	16 (1.7%)	159 (2.0%)	45 (1.0%)	390 (2.9%)	**0.102**	0.092
Estrogen receptor status						
Negative	98 (10.3%)	594 (7.4%)	534 (12.4%)	1589 (11.6%)	-0.062	-0.003

Positive	834 (88.1%)	7339 (91.8%)	3738 (86.6%)	11992 (87.7%)	0.072	0.009
Unknown	15 (1.6%)	62 (0.8%)	43 (1.0%)	88 (0.6%)	-0.043	-0.022
Progesterone receptor status						
Negative	225 (23.8%)	1725 (21.6%)	1247 (28.9%)	3752 (27.4%)	-0.061	0.001
Positive	708 (74.8%)	6206 (77.6%)	3025 (70.1%)	9825 (71.9%)	0.068	0.003
Unknown	14 (1.5%)	64 (0.8%)	43 (1.0%)	92 (0.7%)	-0.038	-0.021

Appendix B: Results Quality Indicator 6b

	Yes		No		Before PSS	After PSS
Indicator 6b	**Norway**	**The Netherlands**	**Norway**	**The Netherlands**	**SMD**	**SMD**
	(N=566)	**(N=5870)**	**(N=186)**	**(N=1133)**		
Year of Diagnosis						
2017	273 (48.2%)	2786 (47.5%)	111 (59.7%)	561 (49.5%)	-0.065	**-0.128**
2018	293 (51.8%)	3084 (52.5%)	75 (40.3%)	572 (50.5%)	0.065	**0.128**
Age						
<40	81 (14.3%)	906 (15.4%)	16 (8.6%)	106 (9.4%)	0.045	0.069
40-49	161 (28.4%)	1696 (28.9%)	40 (21.5%)	204 (18.0%)	0.009	-0.009
50-59	151 (26.7%)	1662 (28.3%)	21 (11.3%)	255 (22.5%)	**0.104**	0.040
60-69	113 (20.0%)	1149 (19.6%)	22 (11.8%)	232 (20.5%)	0.045	-0.055
70-79	55 (9.7%)	403 (6.9%)	44 (23.7%)	207 (18.3%)	**-0.143**	-0.063
80+	5 (0.9%)	54 (0.9%)	43 (23.1%)	129 (11.4%)	**-0.183**	0.018
Histological tumor type						
Ductal	417 (73.7%)	4819 (82.1%)	141 (75.8%)	956 (84.4%)	**0.202**	0.097
Lobular	118 (20.8%)	620 (10.6%)	27 (14.5%)	86 (7.6%)	**-0.262**	-0.043
Other	31 (5.5%)	431 (7.3%)	18 (9.7%)	91 (8.0%)	0.037	-0.098
Differentiation grade						
Well differentiated	31 (5.5%)	460 (7.8%)	23 (12.4%)	99 (8.7%)	0.030	0.081
Moderately differentiated	95 (16.8%)	2663 (45.4%)	51 (27.4%)	514 (45.4%)	**0.577**	0.043
Poorly differentiated	50 (8.8%)	2092 (35.6%)	31 (16.7%)	380 (33.5%)	**0.609**	-0.040
Unknown	390 (68.9%)	655 (11.2%)	81 (43.5%)	140 (12.4%)	**-1.254**	-0.060
HER2 status						
Negative	424 (74.9%)	4165 (71.0%)	140 (75.3%)	848 (74.8%)	-0.077	-0.084
Positive	134 (23.7%)	1682 (28.7%)	42 (22.6%)	255 (22.5%)	0.098	0.096
Unknown	8 (1.4%)	23 (0.4%)	4 (2.2%)	30 (2.6%)	-0.078	-0.042
Estrogen receptor status						
Negative	153 (27.0%)	1996 (34.0%)	46 (24.7%)	294 (25.9%)	**0.137**	0.073
Positive	407 (71.9%)	3870 (65.9%)	137 (73.7%)	834 (73.6%)	**-0.113**	-0.069
Unknown	6 (1.1%)	4 (0.1%)	3 (1.6%)	5 (0.4%)	**-0.132**	-0.022
Progesterone receptor status						
Negative	242 (42.8%)	2806 (47.8%)	82 (44.1%)	496 (43.8%)	0.082	0.030
Positive	318 (56.2%)	3057 (52.1%)	101 (54.3%)	632 (55.8%)	-0.061	-0.027
Unknown	6 (1.1%)	7 (0.1%)	3 (1.6%)	5 (0.4%)	**-0.125**	-0.017

Appendix C: Results Quality Indicator 9a

	Yes		**No**		**Before PSS**	**After PSS**
Indicator 9a	**Norway**	**The Netherlands**	**Norway**	**The Netherlands**	**SMD**	**SMD**
	(N=4625)	**(N=27418)**	**(N=404)**	**(N=1388)**		
Year of Diagnosis						
2017	2411 (52.1%)	13876 (50.6%)	189 (46.8%)	726 (52.3%)	-0.020	0.007
2018	2214 (47.9%)	13542 (49.4%)	215 (53.2%)	662 (47.7%)	0.020	-0.007
Age						
<40	178 (3.8%)	1560 (5.7%)	22 (5.4%)	89 (6.4%)	0.081	0.033
40-49	570 (12.3%)	3959 (14.4%)	61 (15.1%)	283 (20.4%)	0.064	-0.009
50-59	1143 (24.7%)	6837 (24.9%)	110 (27.2%)	366 (26.4%)	0.002	-0.007
60-69	1367 (29.6%)	7446 (27.2%)	126 (31.2%)	360 (25.9%)	-0.057	-0.006
70-79	914 (19.8%)	5730 (20.9%)	76 (18.8%)	241 (17.4%)	0.026	0.015
80+	453 (9.8%)	1886 (6.9%)	9 (2.2%)	49 (3.5%)	-0.091	-0.016
Histological tumor type						
Ductal	3643 (78.8%)	21526 (78.5%)	294 (72.8%)	950 (68.4%)	-0.006	-0.006
Lobular	519 (11.2%)	3336 (12.2%)	75 (18.6%)	271 (19.5%)	0.022	0.020
Other	463 (10.0%)	2556 (9.3%)	35 (8.7%)	167 (12.0%)	-0.015	-0.014
Differentiation grade						
Well differentiated	1077 (23.3%)	6418 (23.4%)	68 (16.8%)	259 (18.7%)	0.010	0.008
Moderately differentiated	2048 (44.3%)	13257 (48.4%)	206 (51.0%)	766 (55.2%)	0.077	0.034
Poorly differentiated	1076 (23.3%)	6552 (23.9%)	104 (25.7%)	280 (20.2%)	0.006	0.006
Unknown	424 (9.2%)	1191 (4.3%)	26 (6.4%)	83 (6.0%)	**-0.182**	-0.093
pT						
1	2812 (60.8%)	17608 (64.2%)	241 (59.7%)	803 (57.9%)	0.066	-0.006
2	1142 (24.7%)	6321 (23.1%)	118 (29.2%)	426 (30.7%)	-0.038	-0.008
3	63 (1.4%)	1036 (3.8%)	16 (4.0%)	105 (7.6%)	**0.146**	**0.101**
Unknown	608 (13.1%)	2453 (8.9%)	29 (7.2%)	54 (3.9%)	**-0.129**	-0.033
pN						
0	2991 (64.7%)	18718 (68.3%)	259 (64.1%)	783 (56.4%)	0.065	0.008
1	1013 (21.9%)	6253 (22.8%)	110 (27.2%)	427 (30.8%)	0.020	0.032
2+	174 (3.8%)	1165 (4.2%)	22 (5.4%)	95 (6.8%)	0.024	0.021
Unknown	447 (9.7%)	1282 (4.7%)	13 (3.2%)	83 (6.0%)	**-0.174**	-0.085
HER2 status						
Negative	4008 (86.7%)	23268 (84.9%)	341 (84.4%)	1182 (85.2%)	-0.046	-0.012
Positive	556 (12.0%)	3533 (12.9%)	59 (14.6%)	163 (11.7%)	0.018	-0.010
Unknown	61 (1.3%)	617 (2.3%)	4 (1.0%)	43 (3.1%)	0.075	0.056
Estrogen receptor status						
Negative	639 (13.8%)	4334 (15.8%)	46 (11.4%)	157 (11.3%)	0.056	0.005
Positive	3929 (85.0%)	22902 (83.5%)	353 (87.4%)	1201 (86.5%)	-0.041	0.010
Unknown	57 (1.2%)	182 (0.7%)	5 (1.2%)	30 (2.2%)	-0.050	-0.055

Progesterone receptor status						
Negative	1417 (30.6%)	8455 (30.8%)	116 (28.7%)	360 (25.9%)	0.003	-0.004
Positive	3157 (68.3%)	18768 (68.5%)	282 (69.8%)	998 (71.9%)	0.005	0.014
Unknown	51 (1.1%)	195 (0.7%)	6 (1.5%)	30 (2.2%)	-0.036	-0.047

Appendix D: Results Quality Indicator 9c

	Yes		No		Before PSS	After PSS
Indicator 9c	**Norway**	**The Netherlands**	**Norway**	**The Netherlands**	**SMD**	**SMD**
	(N=250)	**(N=2550)**	**(N=498)**	**(N=4566)**		
Year of Diagnosis						
2017	136 (54.4%)	1271 (49.8%)	277 (55.6%)	2406 (52.7%)	-0.071	-0.087
2018	114 (45.6%)	1279 (50.2%)	221 (44.4%)	2160 (47.3%)	0.071	0.087
Age						
<40	47 (18.8%)	470 (18.4%)	32 (6.4%)	454 (9.9%)	0.075	0.011
40-49	83 (33.2%)	828 (32.5%)	81 (16.3%)	1003 (22.0%)	0.089	-0.027
50-59	87 (34.8%)	851 (33.4%)	157 (31.5%)	1353 (29.6%)	-0.035	-0.007
60-69	33 (13.2%)	401 (15.7%)	228 (45.8%)	1756 (38.5%)	-0.098	0.024
Histological tumor type						
Ductal	186 (74.4%)	1938 (76.0%)	379 (76.1%)	3264 (71.5%)	-0.056	-0.007
Lobular	41 (16.4%)	359 (14.1%)	75 (15.1%)	836 (18.3%)	0.035	0.043
Other	23 (9.2%)	253 (9.9%)	44 (8.8%)	466 (10.2%)	0.039	-0.043
Differentiation grade						
Well differentiated	50 (20.0%)	457 (17.9%)	82 (16.5%)	668 (14.6%)	-0.049	0.023
Moderately differentiated	126 (50.4%)	1260 (49.4%)	240 (48.2%)	2300 (50.4%)	0.022	0.069
Poorly differentiated	61 (24.4%)	680 (26.7%)	163 (32.7%)	1306 (28.6%)	-0.045	0.086
Unknown	13 (5.2%)	153 (6.0%)	13 (2.6%)	292 (6.4%)	**0.129**	**-0.381**
pT						
1	168 (67.2%)	1525 (59.8%)	275 (55.2%)	1973 (43.2%)	**-0.203**	-0.010
2	71 (28.4%)	592 (23.2%)	203 (40.8%)	1491 (32.7%)	**-0.157**	-0.006
3	7 (2.8%)	100 (3.9%)	15 (3.0%)	534 (11.7%)	**0.255**	0.040
Unknown	4 (1.6%)	333 (13.1%)	5 (1.0%)	568 (12.4%)	**0.463**	-0.006
pN						
0	175 (70.0%)	1734 (68.0%)	310 (62.2%)	2373 (52.0%)	**-0.147**	0.045
1	65 (26.0%)	686 (26.9%)	153 (30.7%)	1524 (33.4%)	0.042	-0.017
2+	7 (2.8%)	60 (2.4%)	31 (6.2%)	493 (10.8%)	**0.110**	-0.077
Unknown	3 (1.2%)	70 (2.7%)	4 (0.8%)	176 (3.9%)	**0.173**	0.033
HER2 status						
Negative	203 (81.2%)	2049 (80.4%)	403 (80.9%)	3757 (82.3%)	0.015	-0.009
Positive	42 (16.8%)	446 (17.5%)	88 (17.7%)	736 (16.1%)	-0.020	0.003
Unknown	5 (2.0%)	55 (2.2%)	7 (1.4%)	73 (1.6%)	0.015	0.018
Estrogen receptor status						
Negative	35 (14.0%)	449 (17.6%)	80 (16.1%)	901 (19.7%)	0.095	**0.102**
Positive	212 (84.8%)	2059 (80.7%)	411 (82.5%)	3622 (79.3%)	-0.089	**-0.104**
Unknown	3 (1.2%)	42 (1.6%)	7 (1.4%)	43 (0.9%)	-0.013	0.015
Progesterone receptor status						
Negative	64 (25.6%)	768 (30.1%)	147 (29.5%)	1590 (34.8%)	**0.107**	0.008
Positive	182 (72.8%)	1739 (68.2%)	345 (69.3%)	2931 (64.2%)	**-0.104**	-0.012

Unknown	4 (1.6%)	43 (1.7%)	6 (1.2%)	45 (1.0%)	-0.009	0.017

Appendix E: Results Quality Indicator 10a

	Yes		No		Before PSS	After PSS
Indicator 10a	**Norway**	**The Netherlands**	**Norway**	**The Netherlands**	**SMD**	**SMD**
	(N=3598)	**(N=16672)**	**(N=150)**	**(N=922)**		
Year of Diagnosis						
2017	1864 (51.8%)	8498 (51.0%)	79 (52.7%)	372 (40.3%)	-0.029	-0.001
2018	1734 (48.2%)	8174 (49.0%)	71 (47.3%)	550 (59.7%)	0.029	0.001
Age						
<40	129 (3.6%)	669 (4.0%)	12 (8.0%)	26 (2.8%)	0.010	0.026
40-49	486 (13.5%)	2247 (13.5%)	14 (9.3%)	48 (5.2%)	-0.009	0.009
50-59	1051 (29.2%)	4632 (27.8%)	36 (24.0%)	98 (10.6%)	-0.047	0.019
60-69	1235 (34.3%)	5194 (31.2%)	30 (20.0%)	160 (17.4%)	-0.071	0.018
70-79	611 (17.0%)	3362 (20.2%)	26 (17.3%)	407 (44.1%)	0.113	-0.042
80+	86 (2.4%)	568 (3.4%)	32 (21.3%)	183 (19.8%)	0.059	-0.045
Histological tumor type						
Ductal	2924 (81.3%)	13624 (81.7%)	117 (78.0%)	739 (80.2%)	0.013	0.010
Lobular	337 (9.4%)	1663 (10.0%)	14 (9.3%)	62 (6.7%)	0.015	-0.029
Other	337 (9.4%)	1385 (8.3%)	19 (12.7%)	121 (13.1%)	-0.033	0.016
Differentiation grade						
Well differentiated	944 (26.2%)	4346 (26.1%)	36 (24.0%)	429 (46.5%)	0.022	-0.019
Moderately differentiated	1705 (47.4%)	8075 (48.4%)	55 (36.7%)	341 (37.0%)	0.018	0.010
Poorly differentiated	849 (23.6%)	3629 (21.8%)	56 (37.3%)	116 (12.6%)	-0.068	0.044
Unknown	100 (2.8%)	622 (3.7%)	3 (2.0%)	36 (3.9%)	**0.056**	-0.085
pT						
1	2684 (74.6%)	12211 (73.2%)	100 (66.7%)	756 (82.0%)	-0.013	-0.030
2	827 (23.0%)	2980 (17.9%)	48 (32.0%)	117 (12.7%)	-0.143	-0.009
3	11 (0.3%)	88 (0.5%)	0 (0%)	5 (0.5%)	**0.037**	-0.019
Unknown	76 (2.1%)	1393 (8.4%)	2 (1.3%)	44 (4.8%)	**0.279**	0.082
pN						
0	2782 (77.3%)	12927 (77.5%)	127 (84.7%)	828 (89.8%)	0.014	0.094
1	727 (20.2%)	3439 (20.6%)	20 (13.3%)	77 (8.4%)	0.001	-0.092
2+	89 (2.5%)	306 (1.8%)	3 (2.0%)	17 (1.8%)	-0.043	-0.016
HER2 status						
Negative	3189 (88.6%)	14428 (86.5%)	131 (87.3%)	845 (91.6%)	-0.054	-0.057
Positive	376 (10.5%)	1982 (11.9%)	16 (10.7%)	49 (5.3%)	0.035	0.043
Unknown	33 (0.9%)	262 (1.6%)	3 (2.0%)	28 (3.0%)	0.061	0.045
Estrogen receptor status						
Negative	376 (10.5%)	2373 (14.2%)	30 (20.0%)	76 (8.2%)	0.094	0.056
Positive	3185 (88.5%)	14242 (85.4%)	117 (78.0%)	842 (91.3%)	-0.070	-0.038
Unknown	37 (1.0%)	57 (0.3%)	3 (2.0%)	4 (0.4%)	-0.086	-0.066
Progesterone receptor status						
Negative	956 (26.6%)	4850 (29.1%)	57 (38.0%)	220 (23.9%)	0.040	0.045

Positive	2607 (72.5%)	11761 (70.5%)	90 (60.0%)	698 (75.7%)	-0.025	-0.034
Unknown	35 (1.0%)	61 (0.4%)	3 (2.0%)	4 (0.4%)	-0.078	-0.060

Appendix F: Propensity Score Stratification with a Federated Learning infrastructure (Personal Health Train)

See pdf.

Propensity Score Stratification with a Federated Learning infrastructure (Personal Health Train)

Dave T. Hamersma[*]

January 2021

[*] Special thanks to Matteo Cellamare for developing the federated GLM at IKNL

1 Introduction

Personal Health Train (PHT) has been introduced by the Netherlands Comprehensive Cancer Organization (IKNL) to answer questions in the field of cancer informatics by incorporating data that are located at different sources. PHT is an open source federated learning infrastructure where the sites using the infrastructure share their statistical model and model parameters instead of sharing sensitive data. The current privacy regulations regarding data exchange, opens possibilities for different approaches of data analysis between countries. Especially in the field of cancer informatics can data exchangeability result in positive outcomes. This, however, comes with new challenges. Due to the systematic differences between countries' populations, an increase in bias by confounding will occur [1]. Confounding is seen as a statistical problem which leads to bias when there are unknown effects contributing to the examined outcome [2, 3]. Rosenbaum et al. [4] proposed the use of propensity scores as a countermeasure to reduce bias by confounding and develop another method for the estimation of an unbiased outcome. The propensity score can be interpreted as the predicted probability of an observation belonging to a group based on their baseline characteristics [4, 5]. In the recent years, the use of propensity scores in large observational studies have been increasing [6]. With the implementation of a propensity score, it is possible to design and analyse observational studies so that it can resemble parts of a randomized control trial [5]. By using this method it is possible to answer questions with data generated from large observational studies, where data from randomized control trials are non-existent or lacking [6]. There are multiple methods based on the propensity score to reduce or eliminate bias by confounding, the most popular being matching, stratification and weighting [7, 6]. However, these methods of propensity score analysis are mainly focused around sharing and merging data to complete the analysis. Within the context of data exchangeability and the current privacy regulations, there is a need of a new approach. The aim of this paper will be exploring the possibilities of implementing a propensity score analysis within a federated learning infrastructure, in particular, Personal Health Train.

2 Propensity Score

The propensity score was originally introduced by Rosenbaum Rubin as a balancing score. Mainly used in the social and health sciences for estimating treatment effects with nonexperimental or observational data [8]. Rosenbaum Rubin proved that observations with the same (or nearest) balancing score, have the same distribution of baseline characteristics. The method is displayed below in formula 1.

$$L(s) = P(X = 1|S = s)$$

Where the propensity score *L(s)* is the probability that the binary treatment X will be chosen by a participant with the baseline characteristics $S = s$.

This states that X and S are independent given the function *L(s)*, which means observations with the same value *L(s)* have somewhat the same distribution of baseline characteristics and are therefore, comparable. The calculation of the propensity score is most often estimated with a binary logistic regression and a logit link function. In the logistic regression the dependent or outcome variable is the binary treatment (e.g. treatment-group or control-group). The independent or conditioning variables are the baseline characteristics. In common practices, the binary logistic regression is calculated with a Generalized Linear Model (GLM).

Generalized linear Model

The term generalized linear model (GLM) refers to a larger class of models popularized by McCullagh and Nelder (1982, 2nd edition 1989). In these models, the response variable y_i is assumed to follow an exponential family distribution with mean μ_i, which is assumed to be some (often nonlinear) function of $x_i^T\beta$. There are three components to any GLM:

- **Random Component** - refers to the probability distribution of the response variable y; e.g. normally distributed in the linear regression, or binomially distributed in the binary logistic regression. More generally, we consider all distribution that can be expressed in the form:

  $$f(y;\theta) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\},$$

  where θ is the canonical parameter, such that $\mathbb{E}(y) = \mu = b'(\theta)$ and $Var(y) = a(\phi)b''(\theta)$. This is also called exponential family. Can be easily showed that, for instance, the canonical parameter for $y \sim N(\mu, \sigma^2)$ is $\theta = \mu$, and the canonical parameter for $y \sim Bin(n, \pi)$ is $\theta = logit(\pi) = log\left(\frac{\pi}{1-\pi}\right)$.

- **Systematic Component** - specifies the explanatory variables $x = (x_1, x_2, \ldots, x_k)$ in the model, more specifically their linear combination define the so called linear predictor

  $$\eta = x^T\beta,$$

  where β must be estimated.

- **Link Function** $g(\cdot)$ - specifies the link between random and systematic components. It says how the expected value of the response relates to the linear predictor of explanatory variables

  $$g(\mu) = \eta$$

  The most commonly used link function for a normal model is $\eta = \mu$, and the most commonly used link function for the binomial model is $\eta = logit(\pi)$. When $\eta = \theta$ we say that the model has a canonical link.

Estimation procedure

In the GLM estimation procedure, the maximum likelihood estimation for β can be carried out via Fisher scoring. The generic $(j+1)$-th step can be calculate by

$$\beta^{(j+1)} = \beta^{(j)} + \Big[-\mathbb{E}l''\big(\beta^{(j)}\big)\Big]^{-1} l'(\beta^{(j)}) \tag{1}$$

where l is the log-likelihood of the entire sample. Ignoring constants, the log-likelihood is

$$l(\theta; y) = \frac{y\theta - b(\theta)}{a(\phi)}$$

After some mathematical operations and using the canonical link $\eta = \theta$, the first derivative and expected second derivative of the log-likelihood are

$$\frac{\delta l}{\delta \beta_j} = \frac{y-\mu}{Var(y)}\left(\frac{\delta\mu}{\delta\eta}\right)x_{ij}$$

$$-\mathbb{E}\left(\frac{\delta^2 l}{\delta\beta_j\delta\beta_k}\right) = \frac{1}{Var(y)}\left(\frac{\delta\mu}{\delta\eta}\right)^2 x_{ij}x_{ik}$$

where x_{ij} (or x_{ik}) is the j-th element of the covariate vector $x_i = x$ for the i-th observation.

It follows that the score vector for the entire data set $y_1, \ldots, y_N$ can be written as

$$\frac{\delta l}{\delta \beta} = X^T A(y-\mu) \tag{2}$$

where $X = (x_1, \ldots, x_N)^T$, and $A = diag\Big[Var(y_i)\big(\frac{\delta\eta_i}{\delta\mu_i}\big)\Big]^{-1}$ and the expected Hessian matrix becomes

$$-\mathbb{E}\left(\frac{\delta^2 l}{\delta\beta_j\delta\beta_k}\right) = X^T W X$$

where $W = diag\Big[Var(y_i)\big(\frac{\delta\eta_i}{\delta\mu_i}\big)^2\Big]^{-1}$.

Therefore the Fisher scoring iteration in 2 can be expressed as

$$\beta^{(j+1)} = \beta^{(j)} + \big(X^T W X\big)^{-1} X^T A(y-\mu) \tag{3}$$

We can arrange the step of Fisher scoring to make it resemble weighted least squares.

Noting that $X\beta = \eta$ and $A = W\frac{\delta\eta}{\delta\mu}$, we can rewrite 2 as

$$\beta^{(j+1)} = \big(X^T W X\big)^{-1} X^T W z \tag{4}$$

where $z = \eta + \frac{\delta\eta}{\delta\mu}(y - \mu)$. Therefore, Fisher scoring can be regarded as Iteratively Reweighted Least Squares (IRWLS) carried out on a transformed version of the response variable.

The IRWLS algorithm can be described as

Algorithm 1 GLM Fisher Scoring algorithm

```
procedure
    initialize $\beta^{(0)}$
        $\eta = X\beta^{(0)}$
        $dev^{(0)}$
    loop
        compute $\mu = g'(\eta)$
            $z = \eta + \frac{y-\mu}{\Delta g'}$
            $W = w\frac{\Delta g'^2}{Var(\mu)}$
        update $\beta^{(j)} = \left(X^T W X\right)^{-1} X^T W z$
            $\eta = X\beta^{(j)}$
        compute $dev^{(j)}$
        if $|dev^{(j)} - dev^{(j-1)}| < \epsilon$ then
            return $\beta^{(j)}$
            end loop
        else
            $j = j + 1$
        end if
    end loop
end procedure
```

where $g(\cdot)$ is the link function, $\Delta g' = \frac{\delta\mu}{\delta\eta}$ is the derivative of the inverse-link function $g'(\cdot)$ with respect to the linear predictor and $w = w_1, \ldots, w_n$ are arbitrary weights assign to the units (by default equal to 1).

The output of the logistic link function is the propensity score L(s). The propensity score is then used in matching, weighing or stratification methods. By using these methods, the effects of confounding can be removed. The methods are explained further below.

- *Matching:* Matching is done based on the propensity scores of the observations that have (almost) the same propensity score. There are many methods of matching, but the most common are *k-nearest neighbour matching* and *exact matching*. With k-nearest neighbour matching an observation in the first group is matched to the closest observation in the other group. In exact matching, the propensity score of the observation in the first group must be exactly the same as the propensity score of the observation in the second group. For both methods applies that if there are no more matches, the unmatched will be discarded.

- *Weighting:* In the case of weighting, all observations will be kept. The idea of weighting is that every observations' propensity score is their respective 'weight'. Their propensity score will be transformed in to weights to be used in a weighted regression.

- *Stratification:* The stratification method uses the propensity score calculated from the binary logistic regression by stratifying the full range of propensity scores in k-strata. The amount of strata is open for debate. It is stated that a five-strata PSS can reduce the bias by at least 90%. By using stratification no observations are discarded.

3 The Federated Propensity Score

Federated learning is a machine learning technique used to create a way of analysing data on decentralised clients without sharing privacy sensible data. Analysis is done separately on each site and, by aggregated statistics, the results are only published and accessible by each site. The importance of federated learning is becoming more apparent as privacy regulations introduces restrictions on data sharing. In the non-federated propensity score the data of two populations/treatments are pooled to calculate the propensity score. Since that is not possible when the data is separated and stored in different locations, the federated propensity score must be calculated in its respective location. After acquiring the propensity score of each observation (which is still in its respective location), these propensity scores are send to the server of *Personal Health Train.* These scores are completely void of privacy, as they represent a predicted outcome of an unknown regression. Now a method of reducing confounding can be applied. After trial and error it became apparent that *stratification* is the best suited for a federated learning infrastructure, as it only requires the complete list of predicted outcomes of an unknown regression.

To further elaborate on the structure of the calculation, the binary logistic regression is explained first:

The main idea behind the federated GLM algorithm is that components of equation 2 can be partially computed in each data sources k and merged together afterwords without pulling together the data.

Let us consider $K \geq 2$ data sources (i.e. cancer registries, schools, banks etc..) and let's denote by n_k the number of observations in the k-th data source such that the total sample size of the study is $n = n_1 + \cdots + n_K$. Furthermore, let us denote by $y_{(k)}$ the n_k-vector of response variable and by $X_{(k)}$ the $(n_k \times p)$-matrix of p covariates for the data source $k = 1, \ldots, K$. It is easy to prove that

$$\begin{aligned} X^T W X &= \left[X_{(1)}^T W_{(1)} X_{(1)}\right] + \cdots + \left[X_{(K)}^T W_{(K)} X_{(K)}\right] \\ X^T W z &= \left[X_{(1)}^T W_{(1)} z_{(1)}\right] + \cdots + \left[X_{(K)}^T W_{(K)} z_{(K)}\right] \end{aligned}$$

where $z_{(K)} = \eta_{(k)} + \frac{y_{(k)} - \mu_{(k)}}{\Delta g'_{(k)}}$ and $W_{(k)} = diag\left[Var\left(y_{(k)}\right)\Delta g'^2_{(K)}\right]^{-1}$.

Therefore, following the structure of algorithm 1, a federated procedure can be described as follows:

Algorithm 2 GLM algorithm

```
Initialization Server
initialize $\beta^{(0)}$
Initialization Node $k$
initialize $\eta_{(k)} = X_{(k)}\beta^{(0)}$
initialize $\mu_{(k)} = g'(\eta_{(k)})$
initialize $dev^{(0)}_{(k)} = f(y_{(k)}\mu_{(k)}, w_{(k)})$

loop
Node $k$
    compute $z_{(k)} = \eta_{(k)} + \frac{y_{(k)} - \mu_{(k)}}{\Delta g'_{(k)}}$
    compute $W_{(k)} = w_{(k)} \frac{\Delta g'^2_{(k)}}{Var(\mu_{(k)})}$
    compute $\mathcal{C}^1_{(k)} = X^T_{(k)} W_{(k)} X_{(k)}$
            $\mathcal{C}^2_{(k)} = X^T_{(k)} W_{(k)} z_{(k)}$
    return to Server $\mathcal{C}^1_{(k)}$ and $\mathcal{C}^2_{(k)}$
Server
    calculate $X^T W X = \sum_{k=1}^{K} C^1_{(k)}$
    calculate $X^T W z = \sum_{k=1}^{K} C^2_{(k)}$
    update $\beta^{(j+1)} = \left(X^T W X\right)^{-1} X^T W z$
    return to Nodes $\beta^{(j+1)}$
Node $k$
    compute $\eta_{(k)} = X_{(k)}\beta^{(j+1)}$
    compute $\mu_{(k)} = g'(\eta_{(k)})$
    calculate $dev^{(j+1)}_{(k)} = f(y_{(k)}\mu_{(k)}, w_{(k)})$
    return to Server $dev^{(j+1)}_{(k)}$
Server
    compute $dev^{(j+1)} = \sum_{k=1}^{K} dev^{(j+1)}_{(k)}$
    if $|dev^{(j+1)} - dev^{(j)}| < \epsilon$ then
        return $\beta^{(j+1)}$
        break loop
    else
        $j = j + 1$
    end if
end loop
```

Now that the regression is calculated, it can be used to predict the response of each observation in every location. The output is then a value between 0 and 1. The next algorithm (full code can be found in appendix A) can be applied:

Algorithm 3 Stratification algorithm

```
    Predicting
 1: Predict the regression on every observation
 2: Assign new column to the dataset with the output of 1
 3: Create numerical output with only the outputs (propensity scores)
 4: Send output to temporary folder in server
    Methods of Trimming (optional)
 5: Non-overlap: Removes propensity scores of Location 1 that are below/higher the lowest/highest propensity score of Location 2 or vice versa.
 6: Percentiles: Removes the $x$ top and bottom percentiles of the each location
 7: Send output to temporary folder in server
    Stratification
 8: Retrieve output from temporary folder
 9: Order the output from minimum to maximum and cut the output in $k$ defined strata
10: Send back the strata output to respective location
11: Paste strata output to propensity score output
```

By now, both datasets acquires a new variable *strata*, which indicates in which stratum every observation is in. Using this information, one can apply **any** calculation within each stratum and calculate the Average Treatment Effect (ATE). For example, if you are interested in the mean age of two countries without confounding, you calculate the mean age in each stratum and then take the mean of the k-strata to get the ATE.

4 Comparing Federated with Non-Federated

In this section, the federated propensity score stratification has been compared to the non-federated version. The data used was gathered from two cancer registries, Cancer Registry Netherlands (CRN) and Norway Cancer Registry (NCR). The data from CRN consists of 32,786 female invasive breast cancer patients diagnosed in hospitals between 2017 to 2018 and the data from NCR included 6377 female invasive breast cancer patients diagnosed between 2017 and 2018. In this case, five breast cancer quality indicators were calculated and compared between the two countries. This means that for every quality indicator a subpopulation is created and defined. The output of a quality indicator is a value between 0 and 100.

The tests were performed on one local computer with R. Propensity Score Stratification in an non-federated manner with base R (glm and quantile of *stats*) and the federated version with *Personal Health Train*. First the output of one subpopulations' logistic binary regression is presented. Secondly, the results of each individual quality indicators (and its subpopulation) are presented.

Listing 1: GLM Non-Federated

```
Call:
glm(formula = formula, family = "binomial", data = in6a)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0358   0.4758   0.6412   0.6892   1.7765

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.31529    0.10296  12.774  < 2e-16 ***
diagyear2018  -0.05565    0.03105  -1.792 0.073075 .
grade2        -0.11580    0.03938  -2.941 0.003277 **
grade3        -0.32434    0.05036  -6.441 1.19e-10 ***
gradeUnknown   0.11070    0.13932   0.795 0.426862
age_bin40-49   0.12745    0.09103   1.400 0.161505
age_bin50-59   0.16836    0.08523   1.975 0.048221 *
age_bin60-69   0.21088    0.08470   2.490 0.012780 *
age_bin70-79   0.52994    0.08711   6.083 1.18e-09 ***
age_bin80+     0.18655    0.09760   1.911 0.055960 .
pT2           -0.16250    0.03747  -4.337 1.44e-05 ***
pT3            0.61167    0.11912   5.135 2.82e-07 ***
pN1            0.10127    0.03977   2.546 0.010882 *
pN2+          -0.29361    0.08490  -3.458 0.000544 ***
pNUnknown      0.94197    0.10446   9.018  < 2e-16 ***
her2Positive  -0.19443    0.05356  -3.630 0.000283 ***
```

```
her2Unknown    1.97564     0.26610    7.424  1.13e-13 ***
erPositive    -0.05130     0.06271   -0.818  0.413337
erUnknown     -1.53961     0.54088   -2.846  0.004420 **
prPositive     0.05799     0.04290    1.352  0.176464
prUnknown     -0.96642     0.54149   -1.785  0.074301 .
histLobulair   0.11602     0.05053    2.296  0.021687 *
histOther     -0.07775     0.05223   -1.488  0.136629
---
Signif. codes:  0   ***   0.001   **   0.01   *
   0.05   .   0.1      1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26603  on 26925  degrees of freedom
Residual deviance: 26077  on 26903  degrees of freedom
AIC: 26123

Number of Fisher Scoring iterations: 5
```

Listing 2: GLM Federated

```
Call:
glm_FL(country ~ diagyear + grade + age_bin + pT + pN +
   her2 + er + pr + hist, family = "binomial")

Coefficients:   Estimate
(Intercept)     1.31529
diagyear2018    -0.05565
grade2          -0.11580
grade3          -0.32434
gradeUnknown    0.11070
age_bin40-49    0.12745
age_bin50-59    0.16836
age_bin60-69    0.21088
age_bin70-79    0.52994
age_bin80+      0.18655
pT2             -0.16250
pT3             0.61167
pN1             0.10127
pN2+            -0.29361
pNUnknown       0.94197
her2Positive    -0.19443
her2Unknown     1.97564
erPositive      -0.05130
erUnknown       -1.53961
prPositive      0.05799
```

```
prUnknown          −0.96642
histLobulair       0.11602
histOther          −0.07775


Degrees of Freedom: 26925 Total (i.e. Null);  26903
    Residual
Null Deviance:        26600
Residual Deviance: 26080          AIC: 1
```

Table 1: Results Cancer Registry Netherlands

QI	Non-Federated	Personal Health Train
1	37 (SD 3.4, CI 34.1-40)	37 (SD 3.4, CI 34.1-40)
2	83.3 (SD 4.8, CI 79.1-87.5)	83.3 (SD 4.8, CI 79.1-87.5)
3	95.2 (SD 0.8, CI 94.5-95.9)	95.2 (SD 0.8, CI 94.5-95.9)
4	36 (SD 6.4, CI 31.3-40.7)	36 (SD 6.4, CI 31.3-40.7)
5	94.9 (SD 4.7, CI 91.8-98)	94.9 (SD 4.7, CI 91.8-98)

Table 2: Results Norway Cancer Registry

QI	Non-Federated	Personal Health Train
1	17.5 (SD 2.5, CI 15.3-19.7)	17.5 (SD 2.5, CI 15.3-19.7)
2	70.8 (SD 5, CI 66.4-75.2)	70.8 (SD 5, CI 66.4-75.2)
3	91.5 (SD 2.7, CI 89.1-93.9)	91.5 (SD 2.7, CI 89.1-93.9)
4	37.4 (SD 10.2, CI 29.8-44.9)	37.4 (SD 10.2, CI 29.8-44.9)
5	95.7 (SD 1.6, CI 94.6-96.7)	95.7 (SD 1.6, CI 94.6-96.7)

5 Conclusion

The Propensity Score Stratification algorithm is working as intended. The non-federated regression model coefficients are the same as the Personal Health Train model, as this was the most federated heavy section, it can be concluded that Personal Health Train is successful in providing a federated learning infrastructure.

References

[1] Nørgaard M, Ehrenstein V, Vandenbroucke JP. Confounding in observational studies based on large health care databases: problems and potential solutions–a primer for the clinician. Clinical epidemiology. 2017;9:185.

[2] Tables S. Statistical methods for research workers. 1925.

[3] Kish L. Some statistical problems in research design. American Sociological Review. 1959:328–338.

[4] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.

[5] Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research. 2011;46(3):399–424.

[6] Yao XI, Wang X, Speicher PJ, Hwang ES, Cheng P, Harpole DH, et al. Reporting and guidelines in propensity score analysis: a systematic review of cancer and cancer surgical studies. JNCI: Journal of the National Cancer Institute. 2017;109(8):djw323.

[7] Zakrison T, Austin P, McCredie V. A systematic review of propensity score methods in the acute care surgery literature: avoiding the pitfalls and proposing a set of reporting guidelines. European Journal of Trauma and Emergency Surgery. 2018;44(3):385–395.

[8] Guo S, Fraser MW. Propensity score analysis: Statistical methods and applications. vol. 11. SAGE publications; 2014.

6 Appendices

In the following appendices the code is presented. The master appendix uses every other appendix.

6.1 Appendix A: Master

```
pss ← function(client, model, stratum, trimming, types){



  USE_VERBOSE_OUTPUT ← getOption('vtg.verbose_output', T)
  lgr::threshold("debug")

  image.name ← "harbor.vantage6.ai/vantage/vtg.pss"

  client$set.task.image(
    image.name,
    task.name ← "PSS"
  )

  # Run in a MASTER container
  if (client$use.master.container) {
    vtg::log$debug(glue::glue("Running␣`pss`␣in␣master␣
        container␣using␣image␣'{image.name}'"))
    # client$use.master.container = F
    # result ← vtg.pss::pss(client, model, stratum,
        trimming, types)
    result ← client$call("pss", model, stratum, trimming,
         types)
    return(result)
  }


  vtg::log$debug("Master:␣Pred")


  #calculate propensity scores and add to the existing
      dataframes
  pr_scores ← client$call("pred", model=model, types=
      types)

  # Apply trimming
  if (trimming ==  'nonoverlap') {

    mins = c()
    maxs = c()
    for (elem in pr_scores) {
      mins = c(mins, min(elem))
      maxs ← c(maxs, max(elem))
    }
    trimming ← c(max(mins), min(maxs))

  }
  pr_scores ← client$call(”trimming”, trimming)

  #calculate combined quantile values
  #done in a master container ~ so this would have to be
      master_q and would go on a file alone
  vtg::log$debug(”Master: Computing quantiles...”)
  # vtg::log$debug(typeof(pr_scores))

  prs = c()
  for (elem in pr_scores) {
    prs ← c(prs, elem)
  }
  q=quantile(prs, seq(0,1,by=1/stratum))
  print(q)
  vtg::log$debug(”Master: Strata”)
  out ← client$call(”strata”, quantiles=q, stratum=
      stratum, types=types)
  return(out)
```

6.2 Appendix B: Propensity Scores predict

```
RPC_pred ← function(df, model, types=NULL){

  vtg::log$debug(”RPC_pred”)

  if(!is.null(types)){
    df=Format_Data(df,types)
  }

  #add pr_score
  pred ← predict(model, newdata=df, type = ’response’)
  df$pr_score=pred

  #df with only pr_scores
  pr_scores=pred

  temp_folder = Sys.getenv(”TEMPORARY_FOLDER”)
  temp_file = file.path(temp_folder, "df.R")
  vtg::log$debug(glue::glue("Writing to {temp_file}"))
  saveRDS(df, file=temp_file)

  vtg::log$debug(paste("pr_scores=", toString(pr_scores))
      )

  return(pr_scores)
}
```

6.3 Appendix C: Trimming

```
RPC_trimming ← function(df, trimming=FALSE) {

    # load dataset from previous set from the temporary
        volume
    vtg::log$debug("RPC_stata: Reading dataframe")
    temp_folder = Sys.getenv("TEMPORARY_FOLDER")
    temp_file = file.path(temp_folder, "df.R")
    df ← readRDS(temp_file)

    vtg::log$debug(glue::glue("trimming = {trimming}"))

    trimmed ← 0
    # legacy trimming
    if (trimming==TRUE){
        vtg::log$debug("BOOL")
        mask ← df$pr_score <= 0.1 | df$pr_score > 0.9
        trimmed ← sum(mask) #summarize amount of trimmed
            observations
        df = df[!(mask),]
    }

    # trimming of nonoverlap
    if ( is.numeric(trimming) == T && length(trimming) ==
        2 ) {
        vtg::log$debug("LIST")
        mask ← df$pr_score <= trimming[1] | df$pr_score >
             trimming[2]
        trimmed ← sum(mask )
        df = df[!(mask),]
    }

    # trimming of percentiles
    if ( is.numeric(trimming) == T && length(trimming) == 1 ){
        vtg::log$debug("VALUE")
        vtg::log$debug(glue::glue("percentile={trimming/100}"))
        mask ← df$pr_score <= (trimming/100) | df$pr_score > (1-trimming/100)
        trimmed ← sum(mask)
        df = df[!(mask),]
    }

    vtg::log$debug(glue::glue("Removed␣{trimmed}␣observations"))

    # write to temporary dataframe
    temp_folder = Sys.getenv("TEMPORARY_FOLDER")
    temp_file = file.path(temp_folder, "filtered_df.R")
    vtg::log$debug(glue::glue("Writing␣to␣{temp_file}"))
    saveRDS(df, file=temp_file)

    return(df$pr_score)

}
```

6.4 Appendix D: Strata

```
RPC_strata ← function(df, quantiles, stratum, types){

  vtg::log$debug("RPC_strata")
  if(!is.null(types)){
    df=Format_Data(df,types)
  }


  # load dataset from previous set from the temporary volume
  vtg::log$debug("RPC_stata:␣Reading␣dataframe")
  temp_folder = Sys.getenv("TEMPORARY_FOLDER")
  temp_file = file.path(temp_folder, "filtered_df.R")
  df ← readRDS(temp_file)

  vtg::log$debug("RPC_stata:␣Computing␣groups")
  df$strata = cut(df$pr_score, breaks = quantiles, labels = 1:stratum, include.lowest = TRUE)
  # write new dataframe (containing the new catergory
     column)
  vtg::log$debug("RPC_stata:␣Writing␣to␣temporary␣
     directory")
  temp_file = file.path(temp_folder, "filtered_df_local.R
     ")
  saveRDS(df, file=temp_file)

  # Some (specific) analysis specific for Dave's master
     thesis
  vtg::log$debug("RPC_stata:␣Specific␣Dave␣analysis")
  res ← matrix(nrow = stratum, ncol = 5)
  x ← 1
  repeat{
    res[x,1] = qualityindicator(df[df$strata == x,],
       variable = "eus6a") #ik pak telkens van de lijst
       out, de aparte dataframes
    res[x,2] = qualityindicator(df[df$strata == x,],
       variable = "eus6b")
    res[x,3] = qualityindicator(df[df$strata == x,],
       variable = "eus9a")
    res[x,4] = qualityindicator(df[df$strata == x,],
       variable = "eus9c")
    res[x,5] = qualityindicator(df[df$strata == x,],
       variable = "eus10a")
    x = x + 1
    if (x > stratum) break
  }

  vtg::log$debug("RPC_stata:␣Reformatting␣results")
  print(res)
  rows = c("eus6a", "eus6b", "eus9a", "eus9c", "eus10a")
  res ← as.data.frame(res)
  colnames(res) ← rows
  row.names(res) ← c(1:stratum)
  print(res)
  #END RESULTS |  AVERAGE TREATMENT EFFECT
  vtg::log$debug("RPC_stata:␣Returning␣results")
  print(colMeans(res))
  return(colMeans(res))


}

qualityindicator ← function(data, variable){
  # (Numerator / Denominator) * 100
  outcome = (sum(data[[variable]] == "Yes") / (sum(data[[
      variable]] == "Yes") + sum(data[[variable]] == "No")
      )*100)
  return(outcome)
}
```