Federated learning has been shown to solve some of the legal issues in multi-centre studies [1–4]. The traditional data analysis method involves collecting all data in one single database, which is not a trivial option with the current GDPR legislation [5,6]. In federated learning, the data stay in each local institution and are not a trivial option with the current GDPR legislation.I n

This iterative model optimisation with distributed data is developed through open-source federated learning. There are some obstacles related to Cox regression analysis since the calculation requires the ranking of survival for all patients across the contributing centres; furthermore, the cohort aggregated hazard at each outcome timepoint needs to be accessible at the central server. This iterative model optimisation with distributed data is
not possible without the risk of the entire data set being reconstructed outside the local institutions, i.e. data leakage [8,9]. Brink et al. have proposed a solution involving a centre-stratified Cox model, where all involved predictive parameters are optimised across the whole decentralised dataset. However, the baseline hazard remains centre-specific, allowing for an analysis of distinct centre differences that can not be explained based on the Cox model [10]. This stratified approach, avoiding the potential data leakage in the model development, was recently used for model validation of a head and neck survival model [11].

Previous publications disagree somewhat on the impact of tumour volume on the overall survival after radiotherapy of larynx cancer. Some studies have shown the volume to be the most significant factor for cancer death in head and neck cancer in general [12,13]. A systematic review on metabolic tumour volume in head and neck cancer by Rijo-Cedeño et al. found that it predicts disease-specific survival and overall survival [14]. Several survival models for larynx cancer have been proposed over the years [15], and some of the older studies indicated that tumour volume had a predictive value in local control after larynx cancer radiotherapy [16–18]. However, some more recent studies have not been able to confirm the predictive power of tumour volume [19–21]. Most of these studies have included relatively few patients, and the tumour volume is often dichotomised or not transformed before analysis which can be problematic due to the skewed volume distribution. However, survival models based on TNM status often fail to separate T3 and T4 patients; thus, there is still interest in investigating if the inclusion of tumour volume might improve the prediction for the higher T classifications.

The presented study aimed to develop a survival model for larynx cancer patients following treatment with radiotherapy with the potential to include tumour volume as an explanatory factor. The model was trained on data from three collaborating centres to increase cohort heterogeneity, using an open-source federated learning platform and a federated stratified Cox algorithm designed to eliminate patient data leakage risks.

Methods and materials

Using an open-source federated learning platform developed in AusCAT (the Australian Computer Assisted Theragnostics network) [22,23], three centres from three different countries provided access to their data from all larynx cancer patients treated with radiotherapy from January 2005 to December 2018.

The inclusion criteria were primary curative radiotherapy and availability of the GTV volume. A maximum of three missing parameters were allowed per patient in the statistical analysis plan (SAP) made before analysis. The CONSORT diagram of the eligible and included patients is shown in supplementary Fig. 2.

Model parameters

Parameters considered to be potentially clinically important factors were: patient age, tumour volume, haemoglobin level, T-classification (T1-T4), N-classification (N0 and N + ), sex, tumour location (glottic and non-glottic), ECOG performance status (P0, P1 and P2 + ), current smoker, and the prescribed biological equivalent dose in fractions of 2 Gy corrected for overall expected treatment time at the start of radiotherapy (EQD₂T) [15].

GTV volume

The gross tumour volume (GTV) was defined as the total contoured GTV, including both GTV primary and GTV node. The tumour volumes were extracted from the clinical treatment plans.

The GTV volume was transformed with a log function to avoid large tumour volumes dominating the regression. Supplementary Fig. 3 shows a histogram over the GTV volume and the transformed log GTV volume. Due to the lower bounded and very skewed distribution of GTV, the use of log-GTV is expected to be a stronger predictor than the raw GTV volume.
Data extraction

At Odense University Hospital, Denmark, the record and verify system (Mosaiq) and the DAHANCA (Danish Head and Neck Cancer) database were used as data sources. Death dates were obtained from the Danish national death registry and were updated in June 2021. All data were recorded prospectively but accessed retrospectively for the study. Data access was granted by the Danish Patient Safety Authority (ref. 3–3013–1798/1).

At the Christie NHS Foundation Trust, Manchester, United Kingdom, patient data were extracted from the record and verify system (Mosaiq), the local patient electronic record system, and radiotherapy treatment planning archives (Pinnacle). The last update was performed in January 2022, and patients alive at that date were censored. Death data were linked from central NHS records (Personal Demographics Service) through weekly synchronisation. All research governance requirements and UK information followed standard UK policy (UK ethical approval ref. 17/NW/0060).

At Liverpool and Macarthur Cancer Therapy Centres, Sydney, Australia, unstructured patient data were retrospectively curated into a structured format and data was then extracted in a structured format from Mosaiq. The date of death was obtained from the NSW Registry of Births, Deaths and Marriages until 31/12/2017 and from hospital records until the extraction date (March 2022) at which point alive patients were censored. Project approval was granted by the NSW Population and Health Services Research Ethics Committee (ref. HREC/16/CIPH5/5).

A predefined ontology was used to convert the centre-specific nomenclature and map the parameters to a common standard [7]. Delivered dose and overall treatment time were defined as the intention to treat at the start of radiotherapy. The first radiotherapy date was used as time zero. Survival times beyond 60 months were time censored as this is the regular follow-up time in all three centres.

The patient characteristics are shown in Table 1, the prognostic index is shown in supplementary Fig. 4, and cumulative characteristics can be seen in supplementary Fig. 5 to 9. All missing values were substituted using multiple imputations (followed by one bootstrap per imputation) by chained equation (MICE), where a series of successive regressions on the available data is used to estimate the missing value. This method allows the utilisation of potential correlations between the different predictors during the imputation process. Since parameter correlation can be different among the contributing centres, all imputation was performed based on data at the local centre.

Statistical methods

Multivariable parameter selection was based on cross-validation utilising best subset selection. Thus all combinations of models with the 11 available parameters were tested; in total $2^{11} = 2048$ models. This selection process is preferred as compared to a standard step-wise approach in which there is a tendency for overfitting due to issues with co-linearity and multiple testing [24, page 68].
Parameter selection was performed using 50 bootstraps within the data sets in each local centre and aggregated globally. In each boot, all models were fitted on the in-boot patients (approximately 2/3) and cross-validated on the out-of-boot patients (approximately 1/3 of the patients). The fitting error on the individual model was then calculated as the average cross-validated likelihood of the 50 boots. Due to the cross-validation approach, the negative likelihood will decrease if the parameters add “true” information to the model. However, when further parameters are added that only contribute to overfitting, the average of the negative cross-validated likelihood will start to increase. The model that maximises the average likelihood (minimises the negative likelihood) among all 2048 models is the best performing model and was referred to as the full model. Since numerous models might be very close in cross-validated likelihood, the simplest model with a likelihood within one standard error of the likelihood of the best performing model was also identified. This approach is in line with the “one standard error rule” suggested by James et al. [25, page 214] as a method to select the simplest model that performs almost as well as the full model. This method forces the dose to be included in the model and is investigated to enable comparison with other previously published models; thus, this is a secondary analysis and not the study’s primary endpoint.

Confidence intervals in validation plots are provided as 95 % confidence intervals and calculated for fixed values of the regression parameters obtained during model optimisation where 1000 bootstraps were used.

### Stratified Cox model

The open-source federated learning used the stratified Cox regression model, with the centres as the stratification variable. This model prevents potential data leakage when the model is iteratively fitted to the local data across the three centres [10]. A stratified Cox regression utilises strata-specific baseline hazards; thus, the model produces individual baseline hazards for each contributing centre. For a perfect survival model, the baseline across the contributing centres will be identical. Therefore, potential differences in the baseline hazard between centres indicate differences in patient cohorts, which the Cox model does not explain.

The validity of the proportional hazard assumption of the Cox model was visualised by plotting the log–log of the Kaplan-Meier estimator for each risk group as a function of time. If these groups have similar baseline hazards, the curves will be similar except for a translation along the y-axis.

### Table 1

Patient characteristics.

<table>
<thead>
<tr>
<th>Patients</th>
<th>Odense</th>
<th>%</th>
<th>Christie</th>
<th>%</th>
<th>Liverpool</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>387</td>
<td>81 %</td>
<td>144</td>
<td>76 %</td>
<td>104</td>
<td>87 %</td>
</tr>
<tr>
<td>Female</td>
<td>91</td>
<td>19 %</td>
<td>45</td>
<td>24 %</td>
<td>15</td>
<td>13 %</td>
</tr>
<tr>
<td>T-classification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>190</td>
<td>40 %</td>
<td>15</td>
<td>8 %</td>
<td>21</td>
<td>18 %</td>
</tr>
<tr>
<td>T2</td>
<td>184</td>
<td>38 %</td>
<td>106</td>
<td>56 %</td>
<td>44</td>
<td>37 %</td>
</tr>
<tr>
<td>T3</td>
<td>66</td>
<td>14 %</td>
<td>58</td>
<td>31 %</td>
<td>41</td>
<td>34 %</td>
</tr>
<tr>
<td>T4</td>
<td>38</td>
<td>8 %</td>
<td>8</td>
<td>4 %</td>
<td>13</td>
<td>11 %</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0 %</td>
<td>2</td>
<td>1 %</td>
<td>0</td>
<td>0 %</td>
</tr>
<tr>
<td>N-classification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N0</td>
<td>346</td>
<td>72 %</td>
<td>141</td>
<td>75 %</td>
<td>82</td>
<td>69 %</td>
</tr>
<tr>
<td>N+</td>
<td>132</td>
<td>28 %</td>
<td>45</td>
<td>24 %</td>
<td>37</td>
<td>31 %</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0 %</td>
<td>3</td>
<td>2 %</td>
<td>0</td>
<td>0 %</td>
</tr>
<tr>
<td>Tumour location</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glottic</td>
<td>267</td>
<td>56 %</td>
<td>72</td>
<td>38 %</td>
<td>68</td>
<td>57 %</td>
</tr>
<tr>
<td>Non-glottic</td>
<td>211</td>
<td>44 %</td>
<td>77</td>
<td>41 %</td>
<td>51</td>
<td>43 %</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0 %</td>
<td>40</td>
<td>21 %</td>
<td>0</td>
<td>0 %</td>
</tr>
<tr>
<td>Chemo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>429</td>
<td>90 %</td>
<td>140</td>
<td>74 %</td>
<td>82</td>
<td>69 %</td>
</tr>
<tr>
<td>Yes</td>
<td>49</td>
<td>10 %</td>
<td>49</td>
<td>26 %</td>
<td>37</td>
<td>31 %</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0 %</td>
<td>0</td>
<td>0 %</td>
<td>0</td>
<td>0 %</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>220</td>
<td>46 %</td>
<td>53</td>
<td>28 %</td>
<td>30</td>
<td>25 %</td>
</tr>
<tr>
<td>1</td>
<td>145</td>
<td>30 %</td>
<td>65</td>
<td>34 %</td>
<td>21</td>
<td>18 %</td>
</tr>
<tr>
<td>2+</td>
<td>86</td>
<td>18 %</td>
<td>31</td>
<td>16 %</td>
<td>4</td>
<td>3 %</td>
</tr>
<tr>
<td>missing</td>
<td>27</td>
<td>6 %</td>
<td>40</td>
<td>21 %</td>
<td>64</td>
<td>54 %</td>
</tr>
<tr>
<td>Smoking</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Former/never</td>
<td>231</td>
<td>48 %</td>
<td>81</td>
<td>43 %</td>
<td>56</td>
<td>47 %</td>
</tr>
<tr>
<td>Current</td>
<td>239</td>
<td>50 %</td>
<td>63</td>
<td>33 %</td>
<td>60</td>
<td>50 %</td>
</tr>
<tr>
<td>Missing</td>
<td>8</td>
<td>2 %</td>
<td>45</td>
<td>24 %</td>
<td>3</td>
<td>3 %</td>
</tr>
<tr>
<td>Haemoglobin</td>
<td>8.9 ± 0.9</td>
<td>mmol/l ± 1SD</td>
<td>8.5 ± 0.9</td>
<td>mmol/l ± 1SD</td>
<td>8.7 ± 0.9</td>
<td>mmol/l ± 1SD</td>
</tr>
<tr>
<td>EQD2T</td>
<td>59.5 ± 2.4</td>
<td>Gy ± 1SD</td>
<td>58.6 ± 2.4</td>
<td>Gy ± 1SD</td>
<td>59.5 ± 2.4</td>
<td>Gy ± 1SD</td>
</tr>
<tr>
<td>GTV</td>
<td>3.0 [0.03–170]</td>
<td>cm3 [range]</td>
<td>3.7 [0.14–156]</td>
<td>cm3 [range]</td>
<td>6.8 [0.45–169]</td>
<td>cm3 [range]</td>
</tr>
<tr>
<td>log(GTV/1cm^3)#</td>
<td>1.1 [−3.5−5.1] [range]</td>
<td>1.5 [−0.8−5.1] [range]</td>
<td>1.9 [−0.8−5.1] [range]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

No missing.

Hemoglobin missing for 28 patients at Christie.

* log was the natural logarithm.
Federated learning

The federated learning software handled all data extraction and figure generation, as specified in the pre-analysis statistical analysis plan. The aggregated data bin sizes were defined per centre, which ensures that the analyses can not be manipulated to harvest all local data.

Results

In total, 786 patients were included in the model development out of 1821 potentially eligible patients in the three centres in the relevant period. Missing GTV was the main exclusion criteria for which 927 were removed. 108 patients were removed due to having more than three missing parameters (Supplementary Fig. 2).

The overall survival for all three centres is shown in Fig. 1. The 95% confidence intervals, which all overlap, are shown in shaded colour.

Parameter selection found the best performing model to include the following six parameters, stated in order of impact on the likelihood (largest impact first): log GTV volume, performance status, age at RT start, smoking during RT, haemoglobin and N-classification. The reduced model, defined by the one standard error rule as the model predicting almost as well as the full model, only contained log GTV and performance status. Fig. 2 shows the cross-validated likelihood for all 2048 models (more specifically, the negative cross-validated log-likelihood minus the similar value for the null model). Each model is indicated by a dot that is placed according to the number of parameters included in the model (upper x-axis). The best performing model within a given number of parameters is the one with the lowest value (the red dots). The parameters indicated on the lower x-axis illustrate the order of importance that the parameters enter the models; hence, the best model with two parameters has log GTV and performance status. The best model with three has the age at start RT added.

The hazard ratios are shown in Fig. 3, and the beta-values for the two models can be found in supplementary table 1. As expected, increasing tumour volume and performance status are associated with increased death hazards.

The calibration plots for the three centres are shown in Fig. 4a for both the full and the reduced model. There is an insignificant difference in the calibration plot between the models, but the full model does have a slightly more extensive prediction range (predicted values have a slightly larger range in the figure). Both models follow the identity line well in all plots, indicating that both models match the data in all three centres. For the reduced model, the Harrell C-indices are for Odense, Christie and Liverpool 0.75 [0.71–0.78], 0.65 [0.59–0.71], and 0.69 [0.59–0.77], respectively. The values are slightly higher for the full model with C-index 0.77 [0.74–0.80], 0.67 [0.62–0.73] and 0.71 [0.61–0.80], respectively.

Centre-specific risk groups are formed based on the quartiles of the locally predicted hazards. The two outer quartiles are one risk group each, while the two centre quartiles are one combined risk group. The prognostic index and risk group thresholds are shown in supplementary Fig. 4. The Kaplan-Meier estimate and the Cox model prediction for each risk group are shown in Fig. 4b for both the full and the reduced model. The full model shows a slightly better separation of the risk groups. A visualisation of the validity of the proportional hazard assumption is shown in supplementary figure 10. There is no indication of proportional hazard violation, in particular not for the two centres contributing most patients where the shape of the curves is more smooth due to the larger number of patients.

The baseline hazard for both models at the three centres are shown in Fig. 5. They are all linear and similar in shape across the centres. The Odense baseline tends to cross the baseline for the two other centres during the latter part of the five years. It should be noted that the scale of the baseline hazard is different between the reduced and full model, which is because the absolute hazard is a product of the baseline hazard and the hazard ratios. Thus if the latter is changed, the former also needs to change to keep the absolute hazard similar.

Discussion

This study demonstrates that it is possible to generate a stratified Cox model based on information from multiple international centres using a federated learning approach, without leaking any patient data. Other radiotherapy studies have used federated learning to model logistic regression [1], support vector machine [2], Bayesian network [3] and Cox proportional hazard [26]. The Cox proportional hazard model has a potential data leakage issue, as the central server needs information about the aggregated local hazard for all death times. Implementing the stratified Cox model can avoid data leakage as the local nodes calculate the partial-likelihood internally; thus, there is no need to share data related to specific timepoints [10]. The Matlab implementation of the stratified Cox model can be used in any federated learning software which supports Matlab executable files distribution.

The primary model of overall survival after radiotherapy for laryngeal cancer from the stratified Cox model contains the log-GTV and performance status, resulting in a well-calibrated model (Fig. 3) that separates three risk groups well (Fig. 4). From a clinical perspective, this model would be simple to use. As larynx cancer patients may die from comorbid illnesses, it is logical for performance status to predict OS. Only a few patients with a performance status of 3+ were contained in the data; thus, they were combined with the performance 2+ patients. Therefore, it is not possible from the current study to separate the effect of performance status 3; however, clinically, these patients are rarely given curative treatment, but it would be expected that performance status 3 would have a higher survival hazard.

The T-classification was the weakest predictor; however, this is perhaps unsurprising as it is a parameter related to the tumour size. When the presumably causal parameter of the GTV volume is added, a proxy parameter, like T-classification, will naturally be weakened. As hypothesised, the GTV volume was the main driver of the larynx survival model in both the reduced and the full model. As stated in the introduction, some studies have shown that volume can be used to predict overall survival and disease control in head and neck cancer [27,28]; this was also observed in laryngeal cancer. Pameijer et al. and Lee et al. used a tumour volume larger than 3.5 cm³ to predict lower survival chances in two groups of approximately 40 patients [16,17]. Dziegielewski et al. found tumour volume to be a significant univariate factor in local control prediction in 107 patients from a single centre [29]. However, more recent studies have not confirmed these findings for laryngeal cancer. Janssens et al. investigated the local control and metastasis-free survival in 270 patients using the GTV primary and GTV node. They did not find the volume to be a prognostic factor for local control or metastasis-free survival when analysing the GTV volume as continuous or dichotomised [21]. De Andrade et al. investigated 145 T3 and T4a laryngeal tumours and found a correlation between tumour volume and overall and disease-free survival; however, this was not significant in the multivariable analyses [30]. Timmerman et al. used the volume of the primary GTV as the tumour volume (without any transformation) in 161 patients. They did not find the volume to impact local control and overall survival in advanced larynx cancer [19]. Given these discrepancies, it is likely that the
Fig. 4. top. The calibration plots for the three centres with the observed survival shown in six patient groups at two and five years. Due to censoring, the observed values are measured as the Kaplan-Meier estimate within each of the six groups at the given time point. The reduced model is shown in dark colours and the full model in light colours. The error bars indicate the 95% confidence interval. Fig. 4 bottom. The Kaplan-Meier estimator and Cox model prediction for each risk group of the three centres. The result of the reduced model is presented on the top row, while similar values for the full model are shown on the lower row. The shaded colour indicates the 95% confidence interval of the Kaplan-Meier estimate.
reason that volume is seen as the strongest impacting parameter in the current study performed across three centres is related to the performed log-transformation of the volume before analysis. It might also be worth mentioning that a considerable number of T1 cancers were included in Odense (190 patients), which was not the case in the two other centres. The primary reason for the low number of T1 cancers in the two centres is that the laryngectomy system and were therefore excluded from analysis. Nevertheless, the model seems to perform well in all three centres.

Unfortunately, it was not possible to separate the primary tumour and nodal volumes in the current data, even though it could be expected that their biological radiation response is different[31–33]. This potential difference is, for example, seen in the better survival prognosis of T3N1 than T1N2/N3 disease. For the primary tumour, the contour would contain the visible tumour extent as per definition; however, for the node, the contour would typically include the entire enlarged lymph node. It is unknown whether the entire node harbours macroscopic tumour cells or is a reactive lymph node with few tumour cells. It is assumed that the tumour cell density would be lower in the lymph node than in the primary tumour. The biology of the two types of GTV is likely quite different.

The fact that there is a macroscopic lymph node with tumour cells could indicate that the cancer is an aggressive type and therefore has a lower chance of being cured. This finding is “confirmed” by the full model that contains N-classification. However, in the current study, N-classification adds only limited predictive power to the model that already included the volume of primary and node GTV. This result is somewhat different from some former studies where N-classification plays an important role in survival [34–37]. Still, that effect diminishes when the GTV-volume is included in the model. It was impossible to distinguish the different N + classifications in the current study as there were relatively few patients in some N + groups.

Although the patient cohorts across the three centres are significantly different, the baseline hazard (Fig. 5) indicates that both models compensate for these differences. The baseline hazards are very similar, suggesting that the log-GTV volume and performance status can predict the outcome quite well for all three centres. The similar baseline hazard also suggests that the model is generalisable; however, this suggestion might need external validation to be confirmed.

Model selection was also made with the dose parameter forced into the model to compare with previous studies. The dose models resulted in dose-related β coefficients of – 0.025 and – 0.02 Gy⁻¹ for the reduced and full models, respectively. It seems plausible that the reason why dose (EQD₂) is not selected in the primary model is linked to the homogeneous doses in the data (supplementary figure 6). However, the β coefficients of the dose are nonetheless comparable to those of Egelmeer et al., who reported a regression constant of – 0.035 Gy⁻¹ [15] (extracted from their nomograms). In the same model, they reported haemoglobin as a prognostic parameter, which the full model of this study confirms. However, compared to the MAASTRO model by Egelmeer et al., the contribution is halved from a regression constant of 0.4 to 0.2 (mmol/L)⁻¹ [15].

The federated learning approach allowed for a full analysis to be run with all data stored at the local centres, and as no patient data left the hospitals, no legal data sharing agreements were required. In the future, when the overhead of establishing federated learning at the local hospitals is minimised, executing research projects like the current study will become much easier than the current standard that needs data sharing agreements.

In conclusion, a stratified Cox model has been developed, and performed across three international centres using federated learning. The simple two-parameter model predicts overall survival for curative radiation-treated laryngeal cancer patients in three different cohorts and shows that the models are well-calibrated with high power of discrimination.

Funding

Danish Cancer Society grant, University of Southern Denmark scholarship, Odense University Hospital scholarship and Danish Cancer Research Fund.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
Larynx cancer survival model using federated learning

Acknowledgements

CB acknowledges support from AgeCare (Academy of Geriatric Cancer Research), an international research collaboration based at Odense University Hospital, Denmark. MF acknowledges support from Cancer Institute NSW Early Career Fellowship 2019/ECF004.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2022.09.023.

References