Radiotherapy and Oncology 144 (2020) 189–200

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

Distributed learning on 20 000+ lung cancer patients – The Personal Health Train

Timo M. Deist [a,b,1], Frank J.W.M. Dankers [a,c,1], Priyanka Ojha [d], M. Scott Marshall [d], Tomas Janssen [d], Corinne Faivre-Finn [e], Carlotta Masciocchi [g], Vincenzo Valentini [f,g], Jiazhou Wang [h], Jiayan Chen [h], Zhen Zhang [h], Emiliano Spezi [i,j], Mick Button [j], Joost Jan Nuyttens [k], René Vernhout [k], Johan van Soest [a], Arthur Jochems [b], René Monshouwer [c], Johan Bussink [c], Gareth Price [e,2], Philippe Lambin [b,2], Andre Dekker [a,2,*]

[a] Department of Radiation Oncology (MAASTRO), GROW – School for Oncology and Developmental Biology, Maastricht University Medical Centre+; [b] The D-Lab: Dpt of Precision Medicine, GROW – School for Oncology and Developmental Biology, Maastricht University Medical Centre+; [c] Department of Radiation Oncology, Radboud University Medical Center, Nijmegen; [d] Department of Radiation Oncology, The Netherlands Cancer Institute – Antoni van Leeuwenhoek, Amsterdam, The Netherlands; [e] The University of Manchester, Manchester Academic Health Science Centre, The Christie NHS Foundation Trust, United Kingdom; [f] Università Cattolica del Sacro Cuore; [g] Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy; [h] Department of Radiation Oncology, Fudan University Shanghai Cancer Center, Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China; [i] School of Engineering, Cardiff University; [j] Velindre Cancer Centre, Cardiff, United Kingdom; [k] Department of Radiation Oncology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Received 22 June 2019
Received in revised form 18 November 2019
Accepted 19 November 2019

Keywords:
Lung cancer
Big data
Distributed learning
Federated learning
Machine learning
Survival analysis
Prediction modeling
FAIR data

ABSTRACT

Background and purpose: Access to healthcare data is indispensable for scientific progress and innovation. Sharing healthcare data is time-consuming and notoriously difficult due to privacy and regulatory concerns. The Personal Health Train (PHT) provides a privacy-by-design infrastructure connecting FAIR (Findable, Accessible, Interoperable, Reusable) data sources and allows distributed data analysis and machine learning. Patient data never leaves a healthcare institute.

Materials and methods: Lung cancer patient-specific databases (tumor staging and post-treatment survival information) of oncology departments were translated according to a FAIR data model and stored locally in a graph database. Software was installed locally to enable deployment of distributed machine learning algorithms via a central server. Algorithms (MATLAB, code and documentation publicly available) are patient privacy-preserving as only summary statistics and regression coefficients are exchanged with the central server. A logistic regression model to predict post-treatment two-year survival was trained and evaluated by receiver operating characteristic curves (ROC), root mean square prediction error (RMSE) and calibration plots.

Results: In 4 months, we connected databases with 23 203 patient cases across 8 healthcare institutes in 5 countries (Amsterdam, Cardiff, Maastricht, Manchester, Nijmegen, Rome, Rotterdam, Shanghai) using the PHT. Summary statistics were computed across databases. A distributed logistic regression model predicting post-treatment two-year survival was trained on 14 810 patients treated between 1978 and 2011 and validated on 8 393 patients treated between 2012 and 2015.

Conclusion: The PHT infrastructure demonstrably overcomes patient privacy barriers to healthcare data sharing and enables fast data analyses across multiple institutes from different countries with different regulatory regimens. This infrastructure promotes global evidence-based medicine while prioritizing patient privacy.

Many current innovations in medicine, including personalized medicine, artificial intelligence, and decision support systems, rely on the sharing of data across healthcare providers. Conventional data analysis requires sharing and centralization of data to answer research questions. However, data sharing is hampered by administrative, political, ethical, and technical barriers [1]. This limits the amount of healthcare data available for life sciences in general as well as for other secondary uses such as healthcare quality assurance.

* Corresponding author.
E-mail address: andre.dekker@maastro.nl (A. Dekker).
[1] authors contributed equally.
[2] authors contributed equally.

https://doi.org/10.1016/j.radonc.2019.11.019
0167-8140/

Distributed (machine) learning reformulates conventional data analysis algorithms so that data centralization becomes unnecessary. Consequently, data transfer agreements are not needed. Distributed algorithms iteratively analyze separate databases and return the same solution as if data were centralized: essentially sharing research questions and answers between databases instead of data.

We are convinced that only sharing research questions (and answers) between healthcare providers is a better, sustainable approach to medical data analysis, and can unlock orders of magnitude more data without violating privacy. To this end, we have developed an infrastructure (see Fig. 1) called the Personal Health Train [2] (PHT) consisting of

- healthcare sites (“stations”) containing FAIR [3] (Findable, Accessible, Interoperable, Reusable) data,
- technical network connections and legal frameworks (“tracks”),
- statistical learning applications (“trains”).

A global community of likeminded healthcare providers and academic partners called CORAL (Community in Oncology for RApid Learning) was initiated at the 2016 European Society for Radiotherapy and Oncology (ESTRO) conference. In various research projects across the globe, CORAL members have worked on the realization of the PHT.

An infrastructure to bring research questions to the data has been demonstrated to work recently in projects such as euroCAT [4,5], DataSHIELD [6] and OHDSI [7]. However, challenges remain in terms of the number of data subjects, number of data providers, and global coverage.

The aim of this study is to show that the PHT distributed learning infrastructure can be scaled to many thousands of patients, approaching the size of national healthcare registries. Specifically, we set the goal (as registered on clinicaltrials.gov [8]) to machine learn a predictive model for post-treatment two-year survival on more than 20 000 non-small cell lung cancer (NSCLC) patients, in at least five healthcare providers from more than five countries—without any patient data leaving a healthcare provider.

Methods

This study was registered on clinicaltrials.gov [8] (https://www.clinicaltrials.gov/ct2/show/NCT03564457) on 11–06-2018 (first posted date: 20–06-2018, actual study start date: 01–07-2018). Official project invitations were sent to eight sites on 18–06-2018 and two additional sites were contacted later but before the deadline of September 1. Fig. 2 shows the project timeline.

In all participating sites, the project was approved by their institutional review boards (IRBs) or was conform to national information and research governance regulations. Given that the PHT is a privacy-by-design infrastructure where no individual patient data leaves the individual healthcare provider, no researcher has access to the data, data is anonymized or pseudonymized, and given the number of patients involved, internal privacy officers often felt informed consent was neither feasible nor necessary.

Patients

Patient cohorts from routine clinical care databases (sites A-B and D-H) or clinical studies (site C) identified as non-small cell lung cancer patients were included in this study (Table 1). Data elements retrieved were

- diagnosis,
- diagnosis date,
- vital status at last follow-up (alive or dead),
- date of last follow-up after the diagnosis date,

and cancer staging as defined by the American Joint Committee on Cancer [23]:

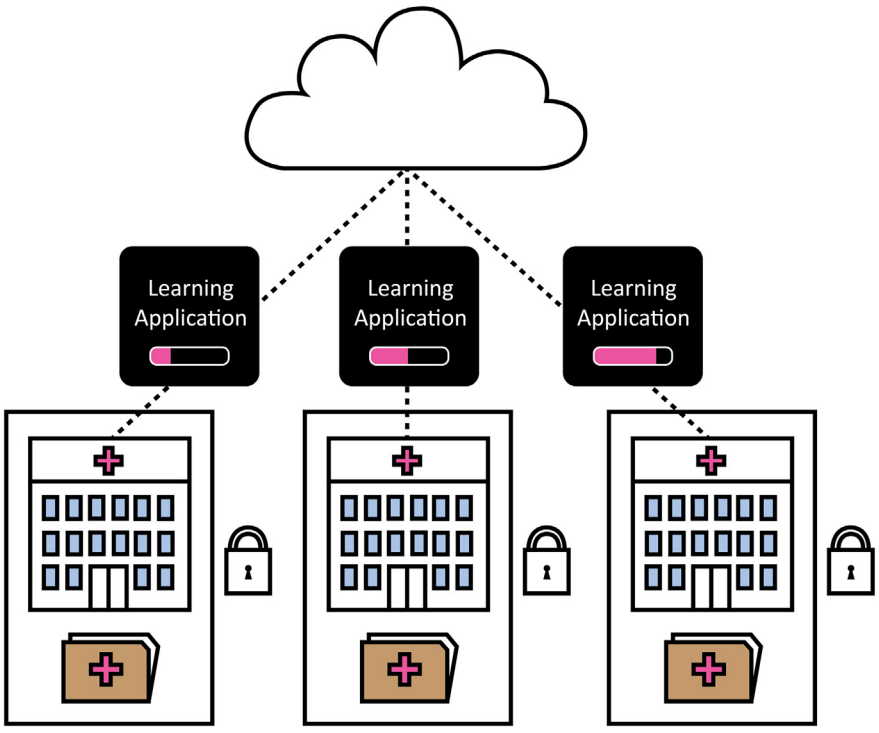

Fig. 1. Personal Health Train infrastructure consisting of a cloud server and network (“tracks”), hospitals with FAIR data (“stations”), and learning applications (“trains”).

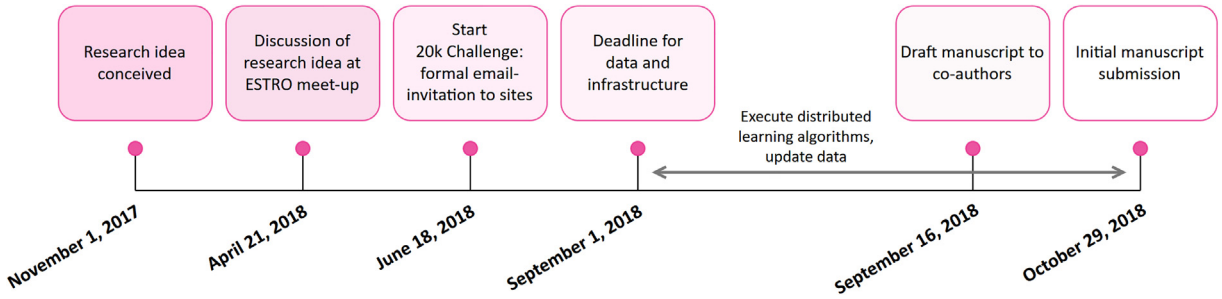

Fig. 2. Project timeline. ESTRO: European Society for Radiotherapy and Oncology.

Table 1
Cohort information. NSCLC: non-small cell lung cancer. SBRT: stereotactic body radiotherapy. RT: radiotherapy. CHART: continuous, hyperfractionated, accelerated radiotherapy.

	Disease	Interval	Treatment
Site A	NSCLC Stage I-IV (histologically confirmed)	January 2008-August 2016	(Chemo-)radiotherapy, surgery, chemotherapy. Filtered for having last follow-up records in 2018 or documented vital status.
Site B	NSCLC, Stage I-IV, histo-cytologically confirmed	October 2004-May 2018	(Chemo-)radiotherapy, chemotherapy, surgery, multimodality treatment.
Site C	NSCLC, 1) Peripheral stage I, 2) stage III	1) 2005–2016, 2) 2008–2013	1) SBRT only, 2) concurrent (chemo-)radiotherapy, surgery.
Site D	NSCLC Stage I-IV (either clinical diagnosis or histologically confirmed)	2004–2017	Definitive radiotherapy (55 Gy in 20 fractions, CHART, concurrent or sequential chemo-radiotherapy or other standard/accepted radical radiotherapy schedules) excluding SBRT or post-surgery adjuvant RT.
Site E	NSCLC, Stage I-IV (either clinical diagnosis or histologically confirmed)	1997–2018	First available T, N, and M staging information of lung cancer patients treated with curative and palliative RT. Includes post-surgery RT, (chemo-)radiotherapy, recurrences.
Site F	NSCLC, Stage I-IV	1982–2018	First available T, N, M, and overall staging information of lung cancer patients treated with curative and palliative RT. Includes post-surgery RT, (chemo-)radiotherapy, recurrences.
Site G	NSCLC, Stage I-IV	1955–2018	First available T, N, M, and overall staging information of lung cancer patients. Includes surgery, (chemo-)radiotherapy.
Site H	NSCLC, Stage I-IV	1971–2018	First available T, N, M, and overall staging information of all lung cancer patients treated with curative and palliative RT. Includes post-surgery RT, (chemo-)radiotherapy, recurrences, SBRT.

- tumor (T) stage,
- lymph node (N) stage,
- metastasis (M) stage,
- overall disease stage.

If the diagnosis date was not available, date of first treatment, date of histology or date of intake were allowed as a surrogate for the date of diagnosis. Various staging editions (AJCC TNM cancer staging editions 1–8) were published and implemented during the period of treatment. Two-year survival was defined as a reported time interval between date of diagnosis and date of last follow-up of more than 2 * 365.24 days with a vital status 'alive' at last follow-up or a reported time interval between date of diagnosis and date of death of more than 2 * 365.24 days. Two-year death was defined as date of death less than 730.48 days after the date of diagnosis. Two-year survival was labelled missing if date of diagnosis, date of last follow-up, or vital status at last follow-up were missing. Two-year survival was also defined as missing if the date of last follow-up was earlier than two years after the date of diagnosis and the vital status at last follow-up was 'alive' (right-censored).

FAIR data model

To make data FAIR, a data model has to be agreed upon between parties. As per prior work [9] we have implemented this model using Semantic Web technology. In Fig. S2, a graphical representation of the model is shown and on github [10] (https://github.com/RadiationOncologyOntology/20kChallenge/wiki) the mapping file containing the full data model including used classes and properties can be found. The 'FAIRness' of our implementation is described in the Supplementary Information (Section I).

FAIR data stations ("stations")

Creating FAIR data out of clinical information systems generally involved the following tools at each institution:

- Source systems: these are the clinical systems in which the data elements required for this study were stored
- Extract, transform, load (ETL): software to extract data from source systems, transform data, and load it into a local data warehouse
- Data warehouse: a local database where data from multiple source systems (within a single institution) are combined
- Mapping: transformation from the local data warehouse schema to medical ontologies, e.g., the Radiation Oncology Ontology [9] (ROO) or the National Cancer Institute thesaurus [11] (NCIt)
- Graph database: Resource description framework (RDF) database where data elements are FAIR

Table 2 shows an overview of the tools used at the various care providers. To support the setup of mapping and graph database software, installation manuals were distributed and remote support was provided. A tutorial describing how to set up software,

Table 2
Overview of tools used to make data FAIR. EMR: electronic medical records.

Provider	Amsterdam (NL)	Cardiff (WAL)	Maastricht (NL)	Manchester (ENG)	Nijmegen (NL)	Rome (IT)	Rotterdam (NL)	Shanghai (CN)
Source systems	NKI-AVL Tumour registry	Canisc (Cancer Network Information System Cymru, NHS Wales Information Services)	HiX (Chipsoft, Netherlands), municipality population registry (survival data)	Clinical Web Portal (in house e-records system). Mosaiq radiotherapy oncology information system. Medway Sigma BI patient administration system.	Radiotherapieweb (in-house EMR), municipality population registry (survival data)	BOA [12] and Speed RO	OpenClinica, Microsoft Access	Chinese EMR
ETL tools	MS SSIS	MATLAB	SAP Business Objects, MATLAB	Pentaho data integration, SQL, Java, Python, R	PHP, SQL, MATLAB	SQL	MATLAB	In-house software
Data warehouse Mapping Graph database	MS SQL Server	MS SQL Server	SAP Business Objects	PostgreSQL D2RQ Blazegraph	SQL Server	SQL Server		None

map data to the required format, and upload it to a local Blazegraph endpoint is available on github [10].

Network for secure application distribution, execution, and communication ("tracks")

For the secure distribution of and messaging between applications, a solution called the Varian Learning Portal (VLP, Varian Medical Systems, Palo Alto, CA) was used. The VLP is a cloud-based system which has implemented user, site, and project management so that a research project consisting of multiple data providers and researchers can securely share applications and communication between applications. To connect the VLP to a local data station, a learning connector is installed at each data provider. The learning connector is a gateway through which applications and communication are handled. The iterative execution of applications and communication between them is called a learning run and each data provider can accept or deny each learning run. All communication and other actions are logged and auditable by members of a given project.

Applications for distributed cohort discovery, and learning ("trains")

The VLP allows a certificate-based upload of applications. Each application group has two parts. One that runs at the VLP in the cloud (master application) and one at each of the sites (site application). Multiple application groups were developed in this project.

- The first application group's aim is cohort discovery. An application is sent to each site to determine and communicate generic statistics (counts) of the available data in the FAIR data station. This cohort discovery application includes a SPARQL Protocol and RDF Query Language (SPARQL) query that can be executed against the graph database. Each site application reports its site statistics to a master application running at the VLP which are then reported back to the researcher who initiated the application. Multiple variations of this application group were employed to generate summary statistics for patient subgroups.
- The second application group aims to train a logistic regression (LR) model. Each LR site application can, given a SPARQL query, train a LR model from the local dataset. The regression coefficients of each site LR model and patient counts are then sent to the master application that reaches consensus in an iterative manner. Fig. 3 illustrates the process followed in the LR application group.
- The third application group validates a given LR model on the sites. An application is sent to each site to compute model performance metrics (RMSE, ROC curve, AUC, calibration plots) and transfers these back to the master application which combines and passes them on to the researcher. Calibration plots reporting calibration-in-the-large and calibration slope are generated following Steyerberg [13] and include Wilson confidence intervals implemented by Winkler and Nichols [14].

The LR model is trained on patients treated between 1978 and 2012 and validated on all patients treated between 2012 and 2015. Only patients with complete diagnosis date, follow-up date, follow-up status, and complete T, N, M, and overall stage after imputation are included. This approach simulates the development of an LR model and sequential validation on new data becoming available over time. This is a TRIPOD type 2b validation [15].

The application used to train the LR coefficients in a distributed manner is based on the Alternating Direction Method of Multipliers (ADMM) and exemplary implementations by Boyd et al. [16,17]. A short description of ADMM is provided in the Supple-

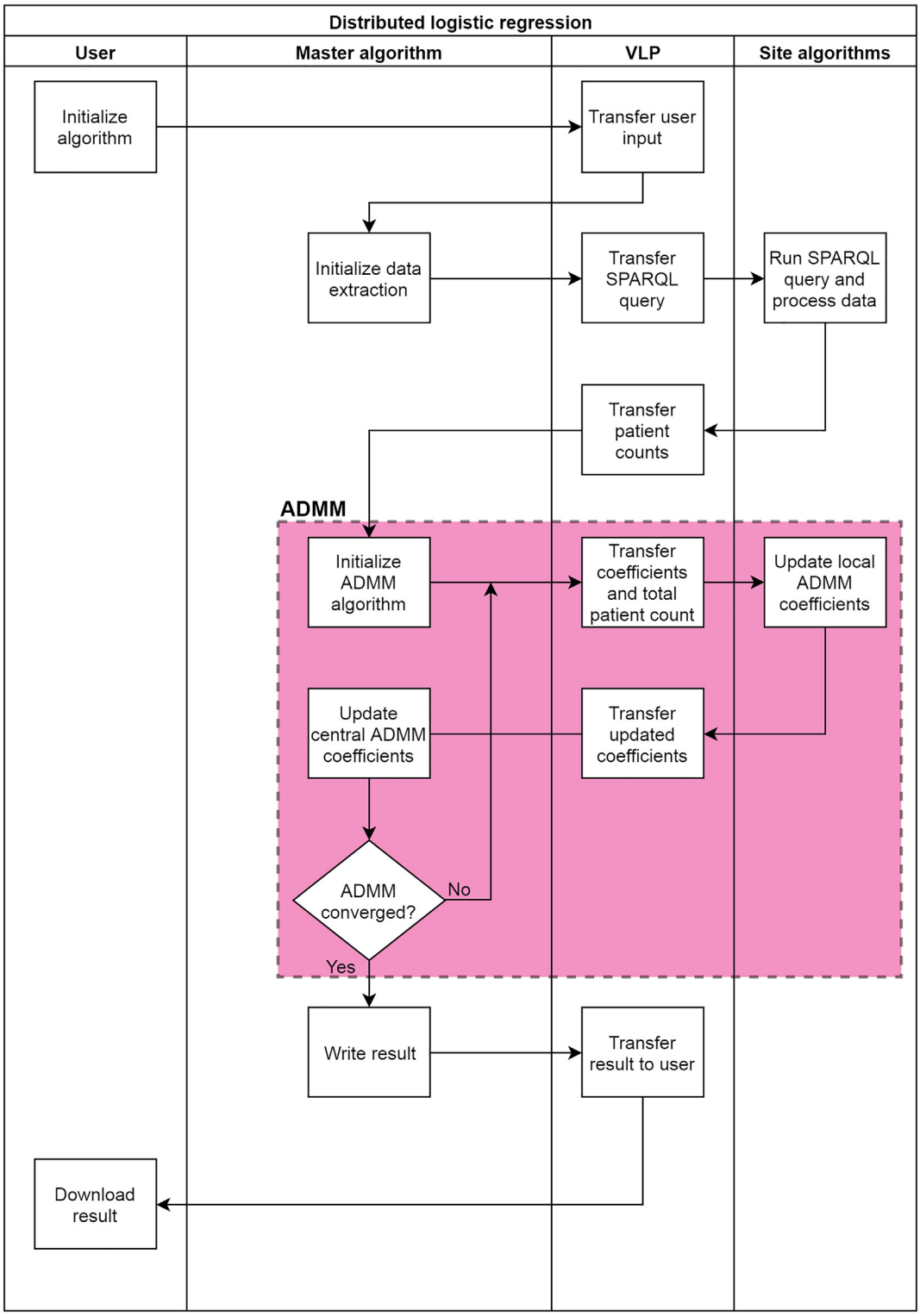

Fig. 3. A simplified process description of the distributed logistic regression application group. VLP: Varian Learning Portal. ADMM: Alternating Direction Method of Multipliers. SPARQL: SPARQL Protocol and RDF Query Language.

mentary Information (Section II). For an excellent technical explanation of ADMM, we suggest Boyd et al. [16]. All application groups are implemented in MATLAB R2018a (Mathworks, Natick, MA). Code and accompanying documentation are available open-source [10] (https://github.com/RadiationOncologyOntology/20kChallenge).

Data processing before LR training

The levels for each variable (T, N, M, and overall stage) are grouped in supercategories (Table 3) to allow regression on data of different AJCC TNM cancer staging editions and to bundle similar categories.

Table 3
Supercategories for T, N, M, and overall stages grouping AJCC TNM cancer staging editions 1–8.

T		N		M		Overall stage	
0	T0	**0**	N0	**0**	M0	**0**	0
1	T1, T1a, T1b, T1c, T1mi, Tis	**1**	N1	**1**	M1, M1a, M1b, M1c	**1**	IA, IA1, IA2, IA3, IB
2	T2, T2a, T2b	**2**	N2	**X**	MX	**II**	IIA, IIB
3	T3	**3**	N3			**III**	III, IIIA, IIIB, IIIC
4	T4	**X**	NX			**IV**	IV, IVA, IVB
X	TX					**Occult**	Occult

T, N, M, and overall stages were dummy-coded to estimate the individual effect of each stage on two-year survival. A reference category was used to avoid multicollinearity issues in the regression model. The combination T1, N0, M0 and overall stage I was chosen as the reference because it is arguably the initial lung cancer stage. For example, the ordinal variable T stage, which takes six values (0 to 4, X), is converted to five binary variables representing T0, T2, T3, T4, TX.

Imputation

If a patient misses entries for one or more of the variables T, N, M, and/or overall staging (but not all of them), imputation of the missing values is attempted. First, the missing values are logically induced from the permitted combinations of T, N, M, and overall stages. If the logical imputation is ambiguous because multiple imputation results are possible, the missing values are imputed probabilistically based on a subset of patients from the training cohort treated at the same site. A detailed imputation process description is presented in Fig. S3 (Supplementary Information) and an outline is given in the Supplementary Information (Section III).

Results

In total, eight healthcare providers (“stations”) were contacted on 18–06-2018 and two additional sites were contacted later. At the deadline of 01–09-2018 (71 days after the first formal project invitation), eight sites (in Amsterdam, Cardiff, Maastricht, Manchester, Nijmegen, Rome, Rotterdam, Shanghai) made NSCLC patient data available in their local database endpoints and two sites did not participate for logistical reasons: delayed response to first formal invitation in one case and too little time to participate after a second round of invitations in another case.

A summary statistics application was sent via the Varian Learning Portal. It computed patient counts for each variable category, displayed in Table 4. Each site confirmed the validity of the summary statistics, a quality control step to ensure that correct data was used for modelling. A total number of 37 090 patients became available in the system. When restricting the search to patients:

- diagnosed or treated from 01-01-1978 (effective date of the AJCC TNM cancer staging edition 1) and before 01–01-2016 (allowing at least two years survival follow-up),
- with complete diagnosis date, follow-up date, and follow-up status (to calculate two-year survival),

the number of available patients decreased to 28 178, which forms the *modelling cohort*. Data of patients diagnosed before 2005 were mainly collected by two sites (with minor contributions from two other sites). Data of patients diagnosed after 2005 were made available by all sites. Overall, recent data was more abundant. More than half of the modelling data was provided by two sites: site G (43.0%) and site E (17.0%). Less than 6% of the modelling data was sourced from three sites: site D (2.4%), site C (2.3%), and site B (1.0%).

Histograms for T, N, M, and overall stage categories after binning into supercategories (Table 3) but before imputation are shown in Fig. 4. Patients with missing or right-censored two-year survival are excluded. The percentage of patients alive at two years differed greatly in the provided data across sites (Fig. 4): from 89.1% in site A to 18.8% in site H. The distribution of T, N, M, and overall stage categories also varied across sites. Notably, T1 clearly dominated in sites A and C but other sites display a more balanced distribution of T categories (Fig. 4a). In sites A-E, N0 is the modal lymph node category but N2 is most frequent in sites F-H (Fig. 4b). All sites report most patients in the M0 category but the decrease in M0 patients correlates loosely with the percentage of patients alive at two years per site, e.g., site H reports 41.4% M1 compared to 8.8% in site A (Fig. 4c). As a direct consequence of the differences in T, N, and M category distributions, the overall stage distribution varies across sites (Fig. 4d).

In general, data completeness is not consistent in the network (Table 4). Sufficient follow-up information to compute two-year survival ranges from 92.1% (site D) to 44.1% (site B). Note that patients with incomplete follow-up (right-censored) have not been included in the modelling cohort displayed in Fig. 4 and Fig. 5. T, N, M, or overall stage information is frequently missing in half of the sites (sites E-H). Overall stage categories are not always reported: sites E and H do not provide overall stage information. Sites G, F, and A miss it for 39.8%, 31.8%, and 2.2% of their patients, respectively.

Based on the temporal distribution of patients in the modelling cohort, we selected patients from 01-01-1978 until and including 31–12-2011 for training and patients from 01-01-2012 until and including 31–12-2015 for validation so that we achieved a split of approximately 2/3 to 1/3 (Fig. 5a).

Only 14 660 patients of 28 178 patients were complete cases (T, N, M, overall stage, and two-year survival) in the modelling cohort (Table 6). Imputation did not result in complete cases for some patients (see methods section for details) yielding a total of 23 203 patients, 14 810 (63.8%) patients for training and 8 393 (36.2%) patients for validation.

The logistic regression application trained a model from the training data (years 1978–2011) with coefficients as displayed in Table 5. The convergence criteria of the algorithm are met after 81 iterations (25 minutes). The convergence of the algorithm is displayed in Fig. 5b: the root mean square error (RMSE, equivalent to the Brier score for binary outcomes) for predicting the probability of two-year survival (left y-axis) in the training cohort decreases per iteration and approaches 0.42. Although the RMSE has stabilized, not all regression coefficients (right y-axis) have converged.

The validation application assessed the model’s performance on the validation cohort (years 2012–2015). The validation performance is described by the combined RMSE for patients from all sites (Fig. 5b), the receiver operating characteristic (ROC) curve per site and their corresponding areas under the curve (AUCs) (Fig. 5c), and by an exemplary calibration plot of the site with most patient data provided for training and validation (site G, Fig. 5d).

Table 4
Summary statistics of all patients provided by the sites. These are patient counts before filtering for the modelling cohort (diagnosed in 1978–2015 with available two-year survival data and at least one stage variable) and before imputation. NSCLC: non-small cell lung cancer.

	Site A	Site B	Site C	Site D	Site E	Site F	Site G	Site H		Site A	Site B	Site C	Site D	Site E	Site F	Site G	Site H
Disease									**Overall stage**								
NSCLC	5214	706	829	785	6211	4110	16,260	2975	Missing	92	3	0	0	6211	1714	7573	2975
T stage									0	208	0	0	0	0	0	1	0
Missing	4	20	0	0	77	807	6703	10	I	0	0	0	0	0	152	282	0
T0	6	1	0	2	3	36	1	16	IA	2413	93	0	141	0	31	704	0
T1	650	30	34	74	322	429	674	200	IA1	0	0	35	0	0	0	6	0
T1a	1694	82	35	42	337	56	351	78	IA2	0	0	191	0	0	0	36	0
T1b	588	40	191	88	285	96	313	117	IA3	0	0	185	0	0	0	31	0
T1c	0	1	185	0	15	16	73	16	IB	501	48	104	141	0	56	373	0
T2	110	75	39	128	1079	803	2138	844	II	0	0	0	0	0	75	101	0
T2a	1032	92	104	139	772	132	472	91	IIA	459	13	49	65	0	17	135	0
T2b	206	18	49	50	194	65	227	45	IIB	188	56	39	78	0	56	235	0
T3	303	165	77	109	1460	523	1936	518	III	0	0	0	2	0	52	621	0
T4	254	151	107	143	1667	1037	1932	639	IIIA	786	187	110	215	0	348	1689	0
TX	164	31	8	10	0	108	1439	396	IIIB	104	103	116	103	0	577	1753	0
Tis	203	0	0	0	0	2	1	5	IIIC	0	0	0	1	0	1	18	0
N stage									IV	199	198	0	39	0	1012	2553	0
Missing	0	20	0	0	14	821	6705	7	IVA	75	0	0	0	0	4	54	0
N0	3649	255	637	384	2756	1041	2830	660	IVB	189	5	0	0	0	15	95	0
N1	520	49	13	153	635	208	598	180	**Diagnosis year**								
N2	777	271	143	215	1835	1132	3510	977	1950–1959	0	0	0	0	0	0	1	0
N3	141	83	36	23	971	810	1437	600	1960–1969	0	0	0	0	0	0	2	0
NX	127	28	0	10	0	98	1180	551	1970–1979	0	0	0	0	0	0	693	1
M stage									1980–1989	0	0	0	0	0	3	2301	362
Missing	2	3	0	0	0	554	6705	4	1990–1999	0	2	0	0	1	16	3192	809
M0	4742	491	829	734	4799	2073	6435	1526	2000–2004	0	5	0	8	1	74	1527	421
M1	87	70	0	8	650	1253	1926	1053	2005	0	18	12	51	223	185	374	83
M1a	92	7	0	11	246	36	164	19	2006	0	15	31	50	313	248	365	78
M1b	285	121	0	20	510	124	497	107	2007	1	24	44	59	276	275	506	68
M1c	1	5	0	0	6	15	107	29	2008	190	123	48	51	314	282	498	95
MX	5	9	0	12	0	55	426	237	2009	214	127	71	42	348	317	528	99
2-year survival									2010	318	92	100	62	401	338	541	125
Missing/Right-censored	614	395	164	62	692	818	3412	477	2011	445	33	117	77	455	306	554	120
No	464	112	258	396	3834	2305	9357	2048	2012	557	32	112	78	626	300	603	121
Yes	4136	199	407	327	1685	987	3491	450	2013	690	34	97	75	692	369	697	100
									2014	971	52	31	70	573	345	755	110
									2015	1057	43	62	63	641	300	763	112
									2016	761	37	103	58	666	302	744	136
									2017	0	35	1	41	562	308	607	112
									2018	10	11	0	0	118	142	163	23

Calibration plots for all other sites are displayed in Fig. S1 (Supplementary Information). Table 6 summarizes patient counts (available in the system and in the modelling cohort before and after imputation) and model performance per site. The validation RMSE almost-monotonically decreases during optimization on the training cohort. Discriminative performance of the model (as measured by the AUC), varies across sites from 0.85 (site A) to 0.58 (site D). Model calibration in site G is good with a calibration-in-the-large of 0.02 and calibration-slope of 0.75 but calibration varies strongly across sites. For example, site A (Supplementary Information, Fig. S1) displays a calibration-in-the-large of 2.39 and a calibration slope of 1.09.

Discussion

We trained a distributed logistic regression model on 14 810 NSCLC patients and validated it on 8 393 patients from eight sites worldwide, yielding a total of 23 203 patients. While we thus easily exceeded the goal of 20 000 by 16.0%, the eight participating sites originate from only five countries which is one country short of the intended goal.

Applying FAIR principles in this project highlighted the challenges in introducing modern data storage and processing approaches in a clinical research context. Semantic web technology allows concepts and relationships between concepts to be coded which makes data more interpretable – an important FAIR principle. The use of semantic web technology requires expertise that is often not present at healthcare institutes. In this project, we worked closely with all partners to support installations. Future projects would benefit from user-friendly software assisting healthcare institutes in transforming their data according to FAIR principles. Creating such software is the goal of an ongoing research project in CORAL.

We observed heterogeneity in modelled variables (T, N, M, and overall stage) and outcome (two-year survival) between sites. Sites provided different cohort types, either (complete) clinical records of heterogeneous NSCLC cases or study cohorts with narrower inclusion criteria which can explain much of this heterogeneity (Table 1). Specifically, site A had a biased inclusion towards surviving patients (89.1% two-year survival, Fig. 4) and site C provided two study cohorts. For both sites, these biases skewed T, N, M, and overall stage distributions towards lower stages. Even for sites providing data based on their full clinical records, different model variable distributions are not surprising since healthcare providers treat different patient subgroups. For example, data in site F originates from a radiotherapy clinic while the data in site G is provided by a comprehensive cancer care center offering different treatments (surgery, (chemo-)radiotherapy, etc.).

For differences in model outcome (two-year survival), there are multiple (possible) causes. For example, site A experienced a biased collection of survival information due to its unavailability

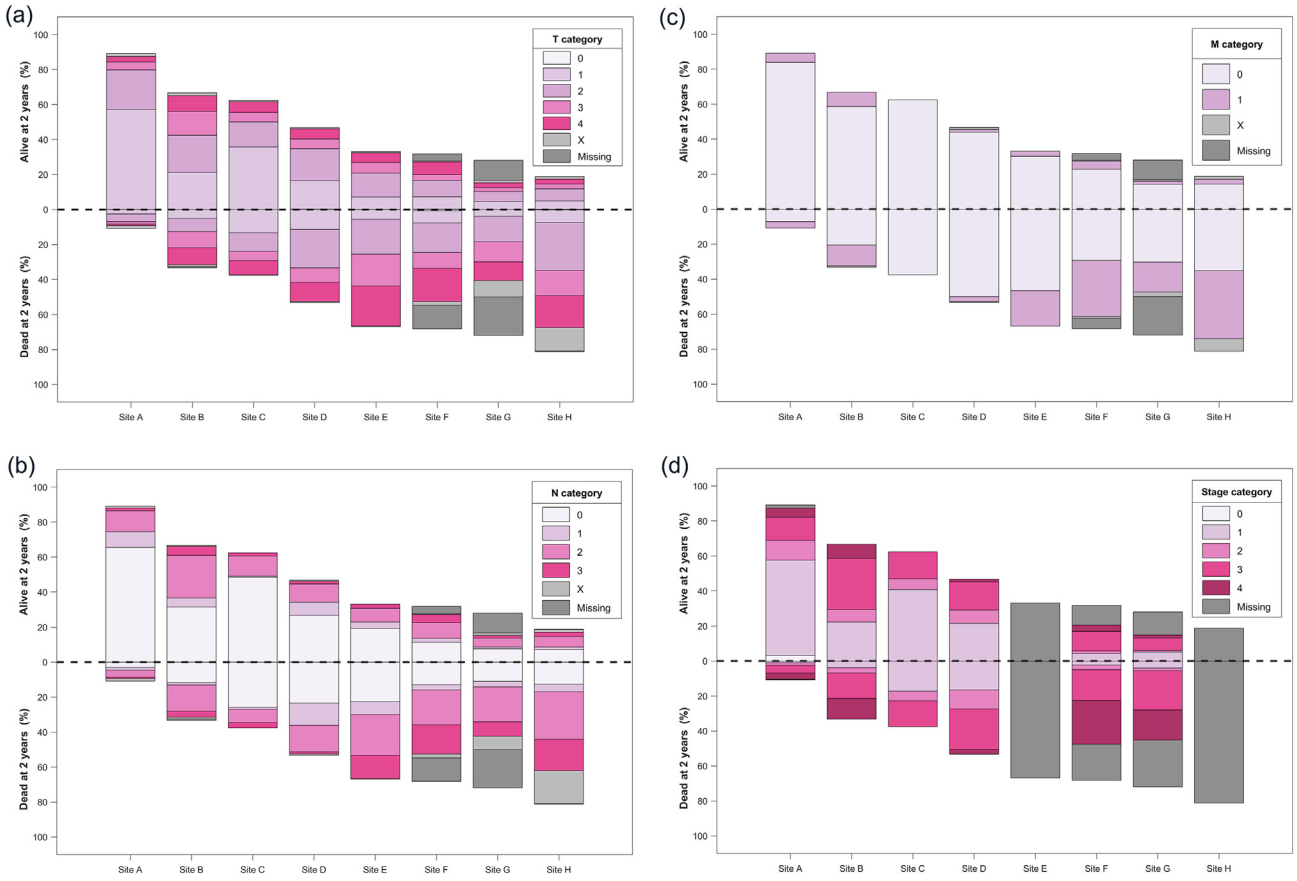

Fig. 4. Distributions of T, N, M, and overall stage supercategories (a, b, c, d, respectively) for patients available for training or validation per site (i.e. the modelling cohort) before selecting for complete cases and imputation. Patients with missing or right-censored two-year survival are excluded. The histograms are separate per site (x-axis) and split for patients alive and dead at two years after diagnosis (above and below x-axis). Patient counts are normalized per site. The vertical position of the entire bar indicates the two-year survival ratio of each site.

in the healthcare provider's Electronic Medical Records (EMR) and the difficulty of retrospectively gathering this missing information when there is no access to survival registries. Furthermore, some sites contributed historical data dating back to 1978 where treatment outcomes were generally worse. Additionally, treatment choices for patient subgroups differ due to national and local treatment guidelines, e.g., patients with metastasized NSCLC.

Heterogeneity throughout the network is generally advantageous for prediction modelling as it allows models to be trained that are generalizable to a wider range of patients. On the other hand, if the difference in cohorts is caused by characteristics not considered by the model, e.g., difference in treatments or data collection biases, then these differences can have a negative effect on model performance. In our study, site A suffered from a biased inclusion of surviving patients. The effect on the trained model should be low as site A only contributed 7.3% of the training cohort (Fig. 5a). However, the usefulness of this dataset for model validation is limited because the performance of this model has not been evaluated for the entire patient population of the site but only for the subgroup following the biased collection (long survivors or recent patients, Table 1). A further inclusion bias is present in site C which provided two study cohorts (predominantly overall stage I and III) for training and validation. Care has to be taken when interpreting validation results: one can only draw conclusions for the patient subpopulation from which the validation dataset has been sampled.

Inter-comparison of summary statistics between sites highlights significant differences in variable distributions that can then be investigated to assure data quality. For example, earlier in this study, the N stage statistics showed one site to have an excess of N3 incidence as compared to other sites. This was subsequently investigated and uncovered a processing error at that site. This role will become increasingly important as outcome modelling studies move away from curated clinical trial datasets and towards routinely collected data and structured information retrospectively extracted from clinical notes.

We also observed varying model performance between sites: the validation cohort AUCs ranged from 0.58 (site D) to 0.85 (site A) and calibration plots (Supplementary Information, Fig. S1) display obvious differences. Multiple factors might influence stable performance across sites: e.g., the aforementioned heterogeneity due to unobserved but important variables, or different staging practices across sites.

We observe that our results are qualitatively in accordance with the AJCC TNM cancer staging system: the regression coefficients of the presented model (Table 5) indicate decreased survival probabilities for increases in T, N, M, and overall stage supercategories (with exception of T4). For example, the regression coefficients for overall stage supercategories decrease from 1.05 for overall stage category 0 to −0.82 for overall stage category IV. Additionally, we quantitatively compared the presented model to the AJCC TNM cancer staging system: we retrieved two-year survival probabilities for the overall stages IA, IB, IIA, IIB, IIIA, IIIB, IV of the AJCC TNM cancer staging edition 7[18] (which is the effective edition of the validation cohort) and predicted two-year survival in the validation cohort. Patients with overall stages other than IA, IB, IIA, IIB,

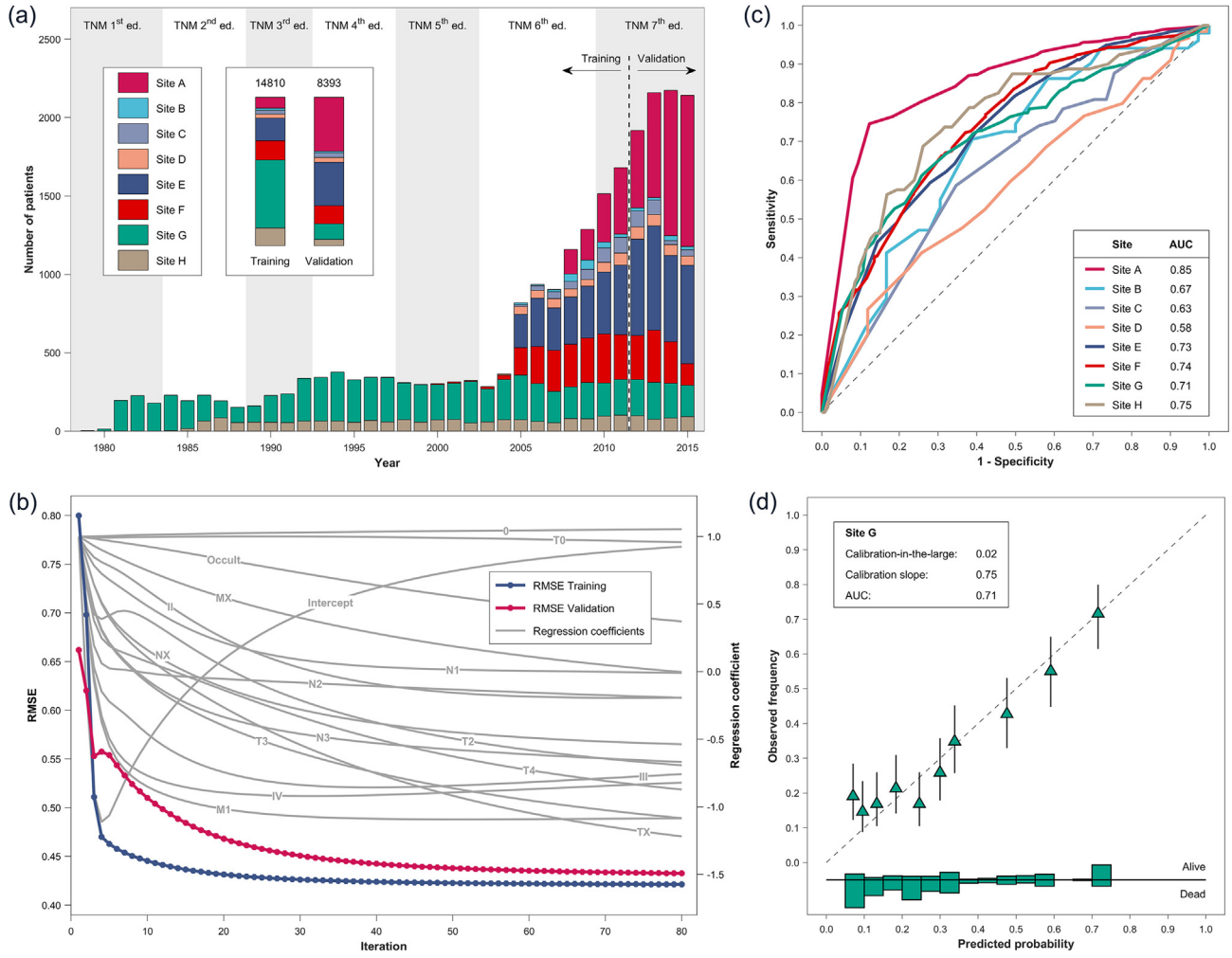

Fig. 5. (a) The number of patients available for training or validation per year per site. (b) Left axis: root mean square error (RMSE) of logistic regression models optimized on the training cohort at a given iteration for training and validation cohorts. Right axis: regression coefficients for T, N, M, and overall stage categories computed by the ADMM algorithm at a given iteration. (c) Receiver operating characteristic curves with area under the curve (AUC) values for the validation cohort per site. (d) Calibration plot of the validation cohort for site G, the site with most training and validation data. Calibration plots for the remaining sites are displayed in Fig. S1.

Table 5
Logistic regression coefficients per supercategory. T1N0M0 and overall stage category I is the reference category.

Intercept	T		N		M		Overall stage	
0.93	**0**	0.96	**0**	ref.	**0**	ref.	**0**	1.05
	1	ref.	**1**	−0.01	**1**	−1.09	**1**	ref.
	2	−0.69	**2**	−0.19	**X**	0.00	**II**	−0.19
	3	−1.08	**3**	−0.67			**III**	−0.76
	4	−0.87	**X**	−0.54			**IV**	−0.82
	X	−1.22					**Occult**	0.37

IIIA, IIIB, IV were excluded because these stages are either not defined or survival probabilities are not reported in TNM edition 7. AUCs of the presented model and the AJCC TNM cancer staging edition 7 coincided (Supplementary Information, Table S2). A discussion of other survival prediction models is available in the Supplementary Information (Section IV).

The results demonstrate the capabilities of distributed learning infrastructures to proffer patient cohorts for statistical analysis. However, it shall be clearly stated that the presented model (Table 5) should not be applied in the clinic as this was not the goal of this study. The modelling methodology could be improved by explicitly encoding different AJCC staging editions and years of treatment which would consider improvements in treatments and outcomes over four decades. Additionally, employing Cox regression instead of logistic regression would allow including right-censored patient data in the analysis.

For this study, we have implemented logistic regression, a tool popular in statistical analysis and machine learning for its simplicity and interpretability. The presented logistic regression algorithm is unpenalized. Penalization might help the individual regression coefficients to converge as it alleviates the multicollinearity problem (Fig. 5b) and will be explored in future studies. We extend the list of distributed methods that are already implemented in the PHT: Bayesian networks [4] and linear support vector machines [5]. Cox regression, a survival analysis methodology to model more than one time point, has been implemented previously in a dis-

Table 6
Patient counts and model performance per site. Sites E and H are listed as incomplete as neither site published overall staging data (which may be imputed from T, N and M stages). AUC: area under the receiver operating characteristic curve. CI: confidence interval using 1000 bootstraps.

Site	Available patients	Modelling cohort patient counts (complete cases, 1978–2015)						Model performance					
		Before imputation			After imputation			Training		Validation			
		Training	Validation	Total	Training	Validation	Total	AUC	95%-CI	AUC	95%-CI	Calibration-in-the-large	Calibration slope
Site A	5214	1050	3024	4074	1084	3058	4142	0.79	[0.75, 0.82]	0.85	[0.83, 0.87]	2.39	1.09
Site B	706	203	87	290	204	87	291	0.71	[0.62, 0.77]	0.67	[0.54, 0.78]	1.04	0.62
Site C	829	390	260	650	390	260	650	0.62	[0.57, 0.67]	0.63	[0.57, 0.69]	0.36	0.59
Site D	785	398	276	674	398	276	674	0.61	[0.55, 0.66]	0.58	[0.51, 0.64]	0.07	0.40
Site E	6211	0	0	0	2265	2458	4723	0.70	[0.68, 0.72]	0.73	[0.70, 0.75]	−0.09	0.85
Site F	4110	1165	520	1685	1906	1017	2923	0.73	[0.71, 0.76]	0.74	[0.71, 0.77]	0.20	0.96
Site G	16,260	6414	873	7287	6803	889	7692	0.74	[0.73, 0.75]	0.71	[0.68, 0.75]	0.02	0.75
Site H	2975	0	0	0	1760	348	2108	0.74	[0.71, 0.77]	0.75	[0.68, 0.80]	−0.43	0.76
Total	37,090	9620	5040	14,660	14,810	8393	23,203						

tributed setting [19]. Distributed learning approaches for other popular machine learning methods are available for future implementation, e.g., (convolutional) neural networks [20].

An alternative to the PHT is DataSHIELD [21], a mature open-source distributed data analysis and machine learning platform with multiple applications. It is based on the open-source software R and Opal data warehouses. The PHT infrastructure differentiates itself from DataSHIELD in multiple aspects:

- it is not limited to R but is compatible with multiple languages (e.g., Java, MATLAB, C#, Python, R),
- it offers analytical flexibility by not limiting the researcher to a fixed function library (DataSHIELD v4.0 comprises 140 R functions [21]),
- it uses Semantic Web technology to store and query data at sites but also allows relational databases and SQL queries.

The presented PHT study only considers a very limited number of clinical data elements (T, N, M, overall stage, diagnosis year, survival follow-up). Arguably, individual predictions need many more data elements. Additional clinical (e.g., age, comorbidities), biological (e.g., genomics, proteomics), imaging (e.g., screening, radiomics [22]) and treatment sources (e.g., radiotherapy treatment planning) are likely to contain relevant data elements for the prediction of a survival outcome. Furthermore, the two-year survival outcome is not sufficient for clinical decision support: quality-of-life, toxicity and cost are also relevant for a balanced decision to be taken. However, due to the limited number of data elements required for inclusion, we could reach very high inclusion numbers and could show that the methodology of distributed learning scales to these numbers. Although the data quality is improving in routine care, the more data elements a study requires, the less complete datasets will be available. As quality improves, future studies are possible where additional data elements (not only prognostic but also predictive for treatment outcomes) can be included and thus better and more clinically relevant models can be developed using the proposed infrastructure.

The PHT enables machine learning studies on more data: more data is generally preferable over too little data. Combining data from multiple institutes, however, comes with challenges faced by any multi-institutional machine learning study (regardless whether it was conducted via a distributed infrastructure or in data centralization projects). Model performance can vary across cohorts (Table 6) or models trained on individual cohorts may perform better. These and other, unexpected results could have different causes, e.g., unobserved confounding factors or different outcome collection standards. Multi-institutional machine learning studies will require a clear methodology to a priori identify and afterwards report on such causes. Experience from and techniques used for clinical trial designs should form the basis for such methodology.

This project shows distributed learning infrastructures are capable of delivering cohort sizes to rival those available to researchers from national registries. However, distributed approaches such as the PHT, where each institute must only satisfy its local information and research governance requirements, ease the bureaucratic burden of learning from internationally separated pools of patients, particularly between countries with differing information governance regimes. Furthermore, the system is much more flexible and makes including additional data elements into analyses a simple process. If an item is not present in a registry dataset, retrospectively adding this information to previous years is very difficult if not logistically impossible. Lastly, the infrastructure provides a mechanism to expedite the external validation of prognostic and predictive models in cohorts from different countries with different patient demographics, organizational cultures, and treatment regimens.

This study has shown that distributed machine learning using Semantic Web technology can be implemented in a short time frame to answer specific research questions. In future work, we will extend CORAL with more cancer centers and include more data elements noted in routine care (we invite all interested parties to contact the corresponding author). As new patients and data elements become available, we expect that the PHT will enable researchers to rapidly train new prediction models: accelerating the speed at which clinical observations are turned into actionable knowledge.

The Personal Health Train infrastructure was deployed across eight healthcare institutes in five countries in four months. A two-year survival prediction model was trained and validated in more than 20 000 non-small cell lung cancer patients. This infrastructure demonstrably overcomes patient privacy barriers to healthcare data sharing and implements distributed data analysis and machine learning across healthcare providers worldwide.

Conflict of interest

Dr. Lambin reports grants/sponsored research from Oncoradiomics, ptTheragnostic/DNAmito. Dr. Lambin reports Advisor (SAB)/presenter fee from Varian, Oncoradiomics, PTT/DNAmito. Dr. Lambin is inventor of two patents on radiomics, one on mtDNA and two non patentable inventions (softwares), licensed to Oncoradiomics & PTT/DNAmito and has (minority) shares in the company Oncoradiomics.

Dr. Dekker has been a paid consultant and received speaker honoraria for Varian Medical Systems. Dr. Dekker is a founder and former employee of Medical Data Works B.V.

Dr. Jochems has (minority) shares in the company Oncoradiomics.

Mr. van Soest has been a paid consultant and received speaker honoraria for Varian Medical Systems. Mr. van Soest is a founder and employee of Medical Data Works B.V.

This study has been completed using Varian Medical Systems software and with technical support by Varian Medical Systems.

Author contributions

- TD, FD conducted the analysis.
- TD, FD, JvS developed software.
- TD, FD, JvS assisted in the technical implementation across sites.
- TD, FD, JvS, AD managed the data collection & technical implementation in Maastricht.
- FD, RM, JB managed the data collection & technical implementation in Nijmegen.
- PO, SM, TJ managed the data collection & technical implementation in Amsterdam.
- CFF, GP managed the data collection & technical implementation in Manchester.
- CM, VV managed the data collection & technical implementation in Rome.
- JW, JC, ZZ managed the data collection & technical implementation in Shanghai.
- ES, MB managed the data collection & technical implementation in Cardiff.
- JJN, RV managed the data collection & technical implementation in Rotterdam.
- TD, FD, TJ, GP, AD developed the study design.
- AJ, RM, JB, GP, PL, AD supervised the research project.
- All authors contributed in the writing of the manuscript.

Acknowledgements

We would like to thank Wolfgang Wiessler (Varian Medical Systems) for his advice and technical support. Sophie Stovold is acknowledged for her work in developing the Velindre (Cardiff) database. Mieke Basten and Thierry Felkers are acknowledged for their work in developing the Nijmegen database. Els Berenschot-Huijbregts and Andras Zolnay are acknowledged for their work in developing the Rotterdam database. Robbert Hardenberg and Tony van de Velde are acknowledged for their work in developing the Amsterdam database.

We would like to thank the following colleagues of the MDTB:
- Giovanna Mantini, [6,7] Department Radiation Oncology.
- A. Martino, [7] Department Radiation Oncology.
- L. Boldrini, [6,7] Department Radiation Oncology.
- A. Damiani, [7] Department Radiation Oncology.
- S. Margaritora, [6,7] Department of Surgery.
- M.T. Cogedo, [7] Department of Surgery.
- F. Lococo, [7] Department of Surgery.
- A. Farchione [7] Department Radiology.
- G. Rindi, [6,7] Department of Pathology.

We wish to acknowledge technical and financial support from the following organizations: Varian Medical Systems (VLP, SAGE); Netherlands Organisation for Scientific Research (grant n° 10696 DuCAT, BIONIC, VWData, grant n° P14-19 Radiomics STRaTegy); Province of Limburg (LIME); Dutch Cancer Society (TraIT2HealthRI, PROTRAIT); Health-RI; Netherlands Federation of University Medical Centres (Data4LifeSciences). This research is also supported by ERC advanced grant (ERC-ADG-2015, n° 694812), EUROSTARS (DART, DECIDE), the European Program H2020-2015-17 Immuno-SABR - n° 733008, PREDICT - ITN - n° 766276, TRANSCAN Joint Transnational Call 2016 (JTC2016 "CLEARLY"- n° UM 2017-8295), Interreg V-A Euregio Meuse-Rhine ("Euradiomics") and Kankeronderzoekfonds Limburg from the Health Foundation Limburg; Cardiff University Data Innovation Research Institute Seedcorn Fund grant n° 23020-AC23024072/16; Velindre NHS Trust Charitable Funds grant n° 2017/12.

Gareth Price and Corinne Faivre-Finn acknowledge the support of Cancer Research UK via funding to the Cancer Research Manchester Centre [C147/A18083] and [C147/A25254]. Corinne Faivre-Finn is supported by the NiHR Manchester Biomedical Research Centre.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2019.11.019.

References

[1] Sullivan R, Peppercorn J, Sikora K, Zalcberg J, Meropol NJ, Amir E, et al. Delivering affordable cancer care in high-income countries. Lancet Oncol 2011;12:933–80. https://doi.org/10.1016/S1470-2045(11)70141-3.
[2] Personal Health Train. Dutch Techcentre for Life Sciences n.d. https://www.dtls.nl/fair-data/personal-health-train/ (accessed September 12, 2018).
[3] Wilkinson MD, Dumontier M, IjJ Aalbersberg, Appleton G, Axton M, Baak A, et al. Guiding Principles for scientific data management and stewardship. Sci Data 2016. https://doi.org/10.1038/sdata.2016.18.
[4] Jochems A, Deist TM, van Soest J, Eble M, Bulens P, Coucke P, et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept. Radiother Oncol 2016;121:459–67. https://doi.org/10.1016/j.radonc.2016.10.002.
[5] Deist TM, Jochems A, van Soest J, Nalbantov G, Oberije C, Walsh S, et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clinical and Translational Radiation Oncology 2017;4:24–31. https://doi.org/10.1016/j.CTRO.2016.12.004.
[6] Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol 2014;43:1929–44. https://doi.org/10.1093/ije/dyu188.
[7] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Studies in Health Technology and Informatics 2015;216:574–8.
[8] 20K Distributed Learning Challenge - ClinicalTrials.gov n.d. https://clinicaltrials.gov/ct2/show/NCT03564457 (accessed September 12, 2018).
[9] Traverso A, van Soest J, Wee L, Dekker A. The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. Med. Phys. 2018;45(10):e854–62. https://doi.org/10.1002/mp.2018.45.issue-1010.1002/mp.12879.
[10] Deist TM, Dankers FJWM, Ojha P, Scott Marshall M, Janssen T, Faivre-Finn C, et al. Code for: Distributed learning on 20 000+ lung cancer patients – The Personal Health Train. 2018. https://github.com/RadiationOncologyOntology/20kChallenge (accessed September 12, 2018).
[11] NCI Thesaurus n.d. https://ncit.nci.nih.gov/ncitbrowser/ (accessed September 12, 2018).
[12] Tagliaferri L, Kovács G, Autorino R, Budrukkar A, Guinot JL, Hildebrand G, et al. ENT COBRA (Consortium for Brachytherapy Data Analysis): interdisciplinary standardized data collection system for head and neck patients treated with interventional radiotherapy (brachytherapy). J Contemp Brachytherapy 2016;8:336–43. https://doi.org/10.5114/jcb.2016.61958.
[13] Steyerberg EW. Evaluation of performance. In: Steyerberg EW, editor. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, New York, NY: Springer New York; 2009, p. 255–80. doi:10.1007/978-0-387-77244-8_15.
[14] Winkler AM. Confidence intervals for Bernoulli trials. Brainder 2012. https://brainder.org/2012/04/21/confidence-intervals-for-bernoulli-trials/ (accessed October 3, 2018).
[15] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Medicine 2015;13:1. https://doi.org/10.1186/s12916-014-0241-z.
[16] Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning 2011;3:1–122.
[17] Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers n.d. http://web.stanford.edu/~boyd/papers/admm_distr_stats.html (accessed October 3, 2018).
[18] Edge SB, American Joint Committee on Cancer, editors. AJCC cancer staging manual. 7th ed. New York: Springer; 2010.

[19] Lu C-L, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc 2015;22:1212–9. https://doi.org/10.1093/jamia/ocv083.
[20] McMahan HB, Moore E, Ramage D, Hampson S. Arcas BA y. Communication-Efficient Learning of Deep Networks from Decentralized Data. ArXiv:160205629 [Cs] 2016.
[21] Wilson RC, Butters OW, Avraam D, Baker J, Tedds JA, Turner A, et al. DataSHIELD – new directions and dimensions. Data Sci J 2017;16:1–21.
[22] Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5..
[23] Editions of the AJCC Cancer Staging Manual n.d. http://cancerstaging.org/references-tools/deskreferences/Pages/default.aspx (accessed September 12, 2018).