Radiotherapy and Oncology xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original article

Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept

Arthur Jochems [a,*,1], Timo M. Deist [a,b,1], Johan van Soest [a,b], Michael Eble [c], Paul Bulens [d], Philippe Coucke [e], Wim Dries [f], Philippe Lambin [a,b,1], Andre Dekker [a,1]

[a] Department of Radiation Oncology (MAASTRO Clinic), Maastricht; [b] GROW – School for Oncology and Developmental Biology, Maastricht University Medical Centre, The Netherlands; [c] Klinik für Strahlentherapie (University clinic Aachen), Germany; [d] Department of Radiation Oncology (Jessa Hospital), Hasselt, The Netherlands; [e] Departement de Physique Medicale (CHU de Liège), Belgium; and [f] Catharina-Hospital Eindhoven, The Netherlands

ARTICLE INFO

Article history:
Received 24 February 2016
Received in revised form 25 August 2016
Accepted 3 October 2016
Available online xxxx

Keywords:
Bayesian networks
Distributed learning
Privacy preserving data-mining
Dyspnea
Machine learning

ABSTRACT

Purpose: One of the major hurdles in enabling personalized medicine is obtaining sufficient patient data to feed into predictive models. Combining data originating from multiple hospitals is difficult because of ethical, legal, political, and administrative barriers associated with data sharing. In order to avoid these issues, a distributed learning approach can be used. Distributed learning is defined as learning from data without the data leaving the hospital.

Patients and methods: Clinical data from 287 lung cancer patients, treated with curative intent with chemoradiation (CRT) or radiotherapy (RT) alone were collected from and stored in 5 different medical institutes (123 patients at MAASTRO (Netherlands, Dutch), 24 at Jessa (Belgium, Dutch), 34 at Liege (Belgium, Dutch and French), 48 at Aachen (Germany, German) and 58 at Eindhoven (Netherlands, Dutch)).

A Bayesian network model is adapted for distributed learning (watch the animation: http://youtu.be/nQpqMIuHyOk). The model predicts dyspnea, which is a common side effect after radiotherapy treatment of lung cancer.

Results: We show that it is possible to use the distributed learning approach to train a Bayesian network model on patient data originating from multiple hospitals without these data leaving the individual hospital. The AUC of the model is 0.61 (95%CI, 0.51–0.70) on a 5-fold cross-validation and ranges from 0.59 to 0.71 on external validation sets.

Conclusion: Distributed learning can allow the learning of predictive models on data originating from multiple hospitals while avoiding many of the data sharing barriers. Furthermore, the distributed learning approach can be used to extract and employ knowledge from routine patient data from multiple hospitals while being compliant to the various national and European privacy laws.

Learning from large volumes of patient data can greatly increase our capacity to generate and test hypotheses about healthcare [1–3]. To capture and use the knowledge contained in large volumes of patient data, predictive models are essential [1,4,5]. Predictive models can be trained on large volumes of data, of patients who have been treated in the past, to make predictions about survival and side-effects of treatment for a patient that has yet to be treated [6–8].

A number of challenges arise when one attempts to train models from routine care patient data. First, model performance is roughly proportional to the number of patients trained on. Patient data are readily available at different hospitals, but unfortunately, sharing these data between hospitals is hampered by ethical, administrative, legal, and political barriers [9]. If these data cannot be shared, models can only be trained on local data with the consequence that performance is limited and external validation (e.g. reproducibility and generalizability) is not possible. External model validation is a prerequisite when developing high quality

* Corresponding author at: Department of Radiation Oncology, Maastricht University (MAASTRO Clinic), Doctor Tanslaan 12, Maastricht 6229 ET, The Netherlands.
E-mail address: arthur.jochems@maastro.nl (A. Jochems).
[1] Authors have contributed equally to this publication.

http://dx.doi.org/10.1016/j.radonc.2016.10.002
0167-8140/

Please cite this article in press as: Jochems A 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), http://dx.doi.org/10.1016/j.radonc.2016.10.002

(e.g. TRIPOD Type 3 or 4) models [10]. In order to circumvent the hurdles associated with sharing data from multiple hospitals, a distributed learning approach may be used. In this approach, many of the barriers regarding data sharing are solved by sending the model training application to the hospitals. The model is trained at the hospitals and then the model is sent back to a central location, rather than centralizing the data. At the central location, the models trained individually at the hospitals are integrated into a single model. Therefore, patient data never leave the hospital and are obscured to the researcher while data are available to the learning application.

A second problem that arises when using data originating from routine clinical practice is data quality, with missing data being the most common problem. Incomplete patient data require many models (logistic regression, support vector machines etc.) to either estimate the missing values or to leave out the patient altogether before model training and/or validation can take place. This impedes model performance [6]. This problem can be avoided by the use of Bayesian networks. The mathematical foundations of Bayesian network models allow effective handling of missing data. We have previously shown that in datasets that have a lot of missing values, a Bayesian network outperforms a support vector machine in terms of discriminative power [6].

Existing techniques to enable distributed learning exist for a number of machine learning algorithms. Some of these techniques focus on horizontal distribution of data, meaning that each center has the same variables but different subjects [11,12]. Other algorithms focus on vertically distributed data, meaning that the data centers each hold different parts of the data for the same subject [13,14]. For Bayesian networks, algorithms exist for learning both the network structure [15–18] and conditional probability tables (CPTs) [19,20]. However, these existing solutions are either tailored to vertically partitioned data or assume the data contains only binary variables.

In this study we learn a Bayesian network model on horizontally partitioned data located at 5 different hospitals using distributed learning, without data leaving the hospital. Our proposed method enables the usage of discretized variables with an arbitrary number of levels. Furthermore, we propose a method by which discretization of continuous variables can take place in a distributed setting. The model predicts dyspnea, a common side effect after radiotherapy of lung cancer [21,22]. Dyspnea is a symptom of radiation-induced lung toxicity, which is developed in 10%-20% of all lung cancer patients treated with radio(chemo)therapy [23]. Our hypothesis is that clinical and treatment parameters, originating from retrospective clinical data from multiple hospitals, can be used to predict patient outcome above the chance level. Furthermore, we expect that we can achieve a similar performance using a model learned from distributed data as we have previously obtained using a centralized approach [24]. Finally, we expect that the results of the distributed algorithm are a close approximation of the results of the centralized algorithm when learned on the same data.

Patients and methods

Data

Clinical data from 287 lung cancer patients, treated with curative intent with chemoradiation (CRT) or radiotherapy (RT) alone were collected and stored in 5 different medical institutes (123 patients at MAASTRO (Netherlands, Dutch), 24 at Jessa (Belgium, Dutch), 34 at Liege (Belgium, Dutch and French), 48 at Aachen (Germany, German) and 58 at Eindhoven (Netherlands, Dutch)). Although these present only a fraction of the data available at these institutions, this study aims to be a proof of principle for distributed learning for which these limited numbers are sufficient. None of the patients received stereotactic body radiotherapy. Patients were treated for their primary lung tumor and were not diagnosed with another tumor in the 5 years before treatment. Out of these 287 patients, 268 patients had post-treatment dyspnea recorded ($\geqslant$Grade 2 according to the CTCv3.0) and were included in the analysis (123 patients at Maastro, 14 at Jessa, 33 at Liege, 42 at Aachen and 56 at Eindhoven). The patient details can be found in Table 1.

Distributed learning infrastructure

Data were placed in a triplestore in subject-predicate-object manner (e.g. Patient # 1 – has post-therapy dyspnea – Grade 3) [25]. In each hospital, data were extracted from the local data sources and then mapped to codes and stored in the local triplestore using an ontology (NCI Thesaurus) [26]. This mapping to codes results in triples which are independent of language and can be accessed by applications using the same query in each hospital. Data stored in this way are said to be Linked Data [27] or Semantically Interoperable [28].

A number of open source software packages were used to implement the infrastructure at each hospital to enable learning. Data were queried and extracted from the electronical medical records using Pentaho (version 6.0.1.0) [29]. Patient identifiers were stored in a secured separate database and an associated 'deidentified' key was generated for use by the system. Deidentified patient data was stored in a PostgreSQL database (version 1.22.1) before conversion to triples using D2RQ took place [30]. Triples were stored in a Sesame server triplestore (version 2.7.7) [31]. Each hospital had its own triplestore with data installed and these triplestores can be queried using SPARQL [32], the query language of the semantic web. Querying of the portals was mediated by the Varian learning portal, a web portal developed by Varian medical systems (Palo Alto, CA). In the Varian learning portal, the researcher can upload his or her model application for learning. The Varian learning portal transmits the model application and validation results back and forth between the central location and the hospitals (Fig. 2).

Bayesian network

A Bayesian network model was developed to predict dyspnea. The model used tumor location (right lower lobe, right middle lobe, right hilus), lung function tests (forced expiratory volume in 1 s, in %, adjusted for age and gender; measured prior to medication), pretreatment dyspnea, cardiac comorbidity, (Non-hypertension cardiac disorder (at baseline), for which treatment at a cardiology department has been given) and timing of chemo (no chemo, sequential or concurrent) to make predictions. Variable selection was based on an earlier study [24]. The network structure of this model was pre-specified by experts and can be found in Fig. 1.

The model's performance was expressed as the Area Under the Curve (AUC) of the Receiver Operating Characteristic Curve (ROC). The maximum value of the AUC is 1.0; indicating a prediction model that perfectly discriminates patients with and without toxicity. A value of 0.5 indicates that patients are only correctly discriminated in 50% of the cases, e.g., as good as chance.

A Bayesian network is a probabilistic graphical model that represents a set of variables and their dependencies in a directed acyclic graph (DAG). Within the DAG, variables are depicted as nodes and statistical dependencies are represented as directed edges. The joint probability of variables $X_1, \ldots, X_n$, can be decomposed into a product form of conditional probability distributions:

Please cite this article in press as: Jochems A 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), http://dx.doi.org/10.1016/j.radonc.2016.10.002

Table 1
Overview of patient characteristics per hospital.

	Jessa	UK Aachen	Chulg	Eindhoven	Maastro
Number of patients	24	48	34	58	123
Neoplasm					
Lung Carcinoma	10 (42%)	2 (4%)	0 (0%)	1 (2%)	6 (5%)
Non-Small Cell Lung Carcinoma	9 (37%)	30 (63%)	27 (80%)	40 (69%)	91 (74%)
Small Cell Lung Carcinoma	5 (21%)	16 (33%)	7 (20%)	17 (29%)	26 (21%)
Missing	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Dyspnea score (pre-treatment)					
CTCAE Grade < 2	6 (25%)	36 (76%)	20 (59%)	50 (86%)	119 (97%)
CTCAE Grade ⩾ 2	8 (33%)	6 (12%)	13 (38%)	6 (10%)	4 (3%)
Missing	10 (42%)	6 (12%)	1 (3%)	2 (3%)	0 (0%)
Dyspnea score (post-treatment)					
CTCAE Grade < 2	8 (33%)	36 (76%)	20 (59%)	50 (86%)	89 (72%)
CTCAE Grade ⩾ 2	6 (25%)	6 (12%)	13 (38%)	6 (10%)	34 (28%)
Missing	10 (42%)	6 (12%)	1 (3%)	2 (3%)	0 (0%)
Gender					
Female	6 (25%)	14 (29.2%)	13 (38%)	25 (43%)	46 (37%)
Male	18 (75%)	34 (70.8%)	21 (62%)	33 (57%)	77 (63%)
Missing	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Chemo timing					
None	6 (25%)	2 (4%)	0 (0%)	6 (10%)	16 (13%)
Sequential	4 (18%)	4 (8%)	3 (8%)	25 (43%)	22 (18%)
Concurrent	13 (53%)	28 (59%)	31 (91%)	27 (47%)	84 (69%)
Missing	1 (4%)	14 (29%)	0 (0%)	0 (0%)	0 (0%)
FEV1Percentage	82 (±22)	68 (±20)	72 (±22)	81 (±25)	78 (±22)
Missing	4 (16%)	23 (48%)	0 (0%)	21 (37%)	0 (0%)
ECOG performance status					
1	7 (29%)	0 (0%)	9 (26%)	23 (40%)	34 (28%)
2	10 (42%)	0 (0%)	22 (65%)	29 (50%)	68 (55%)
3	1 (4%)	0 (0%)	2 (6%)	6 (10%)	16 (13%)
4	0 (0%)	0 (0%)	0 (0%)	0 (0%)	4 (3%)
5	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1%)
Missing	6 (25%)	48 (100%)	1 (3%)	0 (0%)	0 (0%)

FEV1 Percentage: forced expiratory volume in 1 s, in %, adjusted for age and gender; measured prior to medication. CTCAE: Common Terminology Criteria for Adverse Events. ECOG: Eastern Cooperative Oncology Group.

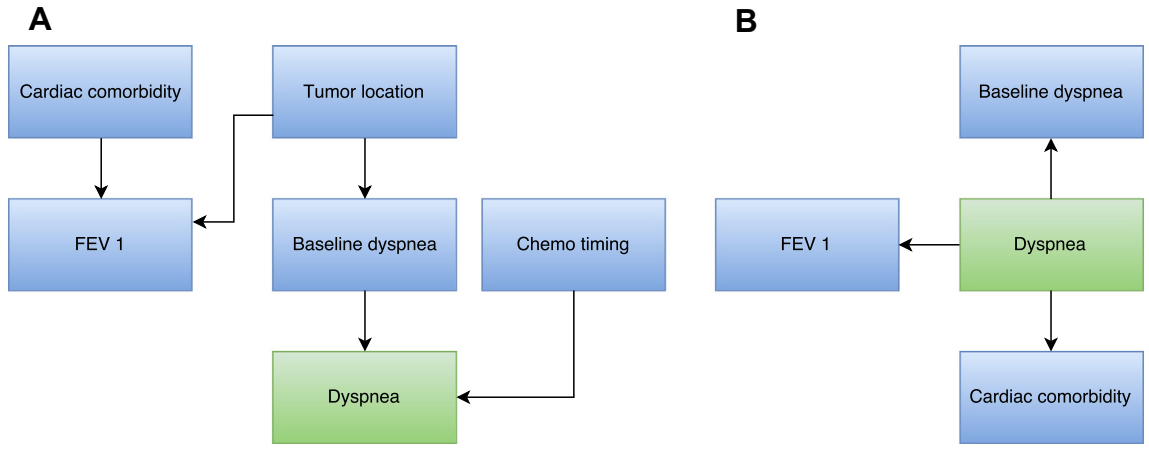

Fig. 1. Network structure of the Bayesian network model. The underlined node, Dyspnea, is the prediction variable. FEV1: forced expiratory volume in 1 s, in %, adjusted for age and gender. (A) Network structure determined by experts. (B) Network structure computed by structure learning algorithm.

$$P(X) = \prod_{i=1}^{n} P(X_i | Pa(X_i))$$

in which $Pa(X_i)$ indicates the parents of X_i in the network. We have determined the DAG dependencies based on expert knowledge. The conditional probability tables associated with each variable have been computed using the expectation maximization (EM) algorithm [33]. All continuous variables were discretized using a method described by Kuschner and colleagues [34].

To evaluate the effect of network structure on model performance, we have compared performance of three different network structures by validating on the Eindhoven dataset. In addition to the expert defined structure, we have used the same structure without chemo timing to investigate the effect of chemotherapy use on dyspnea. To explore the effects of using a data-driven approach to determine the network structure, we have used the path condition (PC) algorithm to determine the network structure [35] based on the largest dataset (Maastro).

Please cite this article in press as: Jochems A 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), http://dx.doi.org/10.1016/j.radonc.2016.10.002

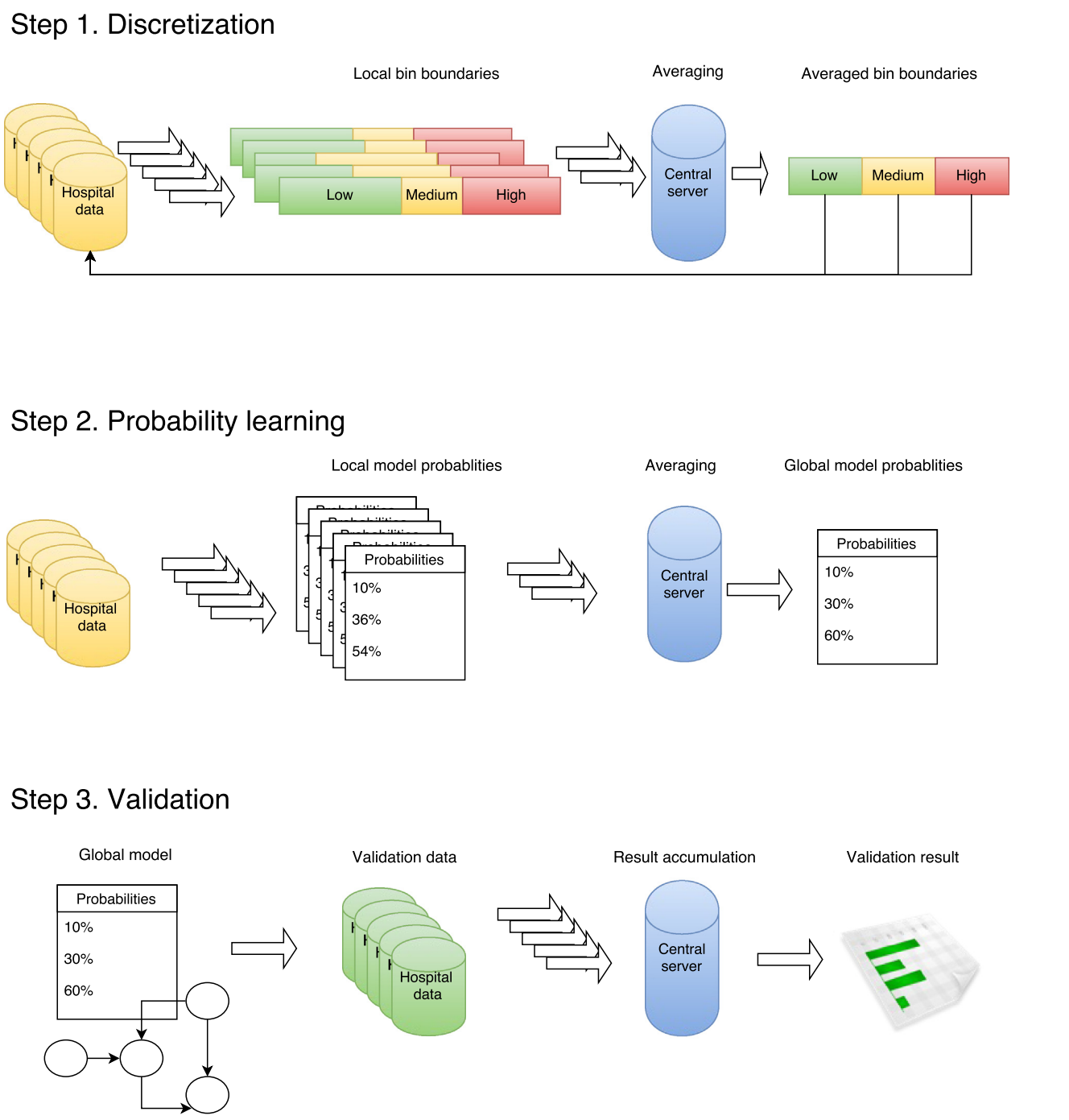

Fig. 2. Schematic overview of the distributed learning algorithm. In the first step, locally found bin boundaries are sent back to the central location. Here, a weighted average is computed on these boundaries, and these are sent back to the sites. In the second step, conditional probability tables are learned at each hospital. These probability tables are sent back to the central location, at which they are averaged in proportion to the number of patients at each hospital. The averaged conditional probability tables are sent back to the sites. In the last step, the model is validated on the data and the predicted probabilities of the model are sent back to the central location.

Distributed learning

In order to realize distributed learning, existing techniques for learning Bayesian networks had to be adapted. A schematic overview of the procedure can be observed in Fig. 2. First, continuous variables were discretized in three bins according to the method described by Kuschner and colleagues [34]. Locally computed bin boundaries were transmitted to a central location. At the central location, the bin boundaries were averaged. Each site contributed to the average in proportion to the number of patients that were located on that site. The averaged bin boundaries were transmitted back to the hospitals. In the second step, the CPTs were obtained by learning locally from each hospital. The CPTs were sent to the central location, where they were combined by weighted averaging. Individual table entries were weighted in proportion to the number of patients available at the hospital. The weighted CPTs, which comprise the global model, were sent back to each site to be validated. In the third step, the model was validated using the data available on the site. The predicted probabilities of the model and corresponding actual outcomes were transmitted back to the central location.

Please cite this article in press as: Jochems A 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), http://dx.doi.org/10.1016/j.radonc.2016.10.002

We conducted a 5-fold cross-validation of the model. Cross-validation was done by algorithmically selecting 80% of all patients at random at each hospital and using this 80% for learning the model (steps 1 and 2 in Fig. 2). The models learned at individual hospitals were combined at a central location. The combined model was subsequently sent back to each hospital to be validated on the remaining 20% of patients on each site. This was repeated 5 times.

In order to assess the overall performance of the technique used, we did an additional experiment with the network structure as described in Fig. 1. In this experiment, we learned the network on all data, except for the data of a single center (steps 1 and 2 in Fig. 2). The model is subsequently validated on that last center, yielding an ROC curve (step 3 in Fig. 2). The learning and validation steps were repeated so the model could be validated on each center.

To evaluate the feasibility of learning high quality models from larger volumes of data and higher numbers of hospitals using the method presented in this study, additional computational analyses were done. First, datasets of 1000, 10,000 and 100,000 patients were created by taking random samples from the data available at MAASTRO clinic. These data were partitioned in 10 to 100 subsets to simulate the different hospitals. The algorithm proposed in this study was used on these partitions and performance of the distributed algorithm model was compared to the model learned locally. Both models were validated on the data used to train the models. Model performance was evaluated using average difference in probabilities in the CPTs and difference in AUC between the distributed model and non-distributed model. To evaluate the performance of the algorithm under varying levels of missing data, the experiment was repeated with 0%, 20% and 40% randomly set missing data elements in the training data.

In all participating hospitals, internal review board approval was obtained.

Statistical analysis

Data preprocessing was done in Matlab (MathWorks, Natick, MA, USA) The Bayesian network model was programed in Java using the open-source jSmile framework of the Dynamic Systems Laboratory of Pittsburg University [36] and made freely available for academic purposes by BayesFusion, LLC (http://www.bayesfusion.com/).

Analysis of ROC curves was done in R, version 3.1.3 (R foundation for statistical computing, Vienna, Austria) using the pROC [37] and cvAUC [38] packages. Comparison of ROC curves and computation of confidence intervals of AUC values was done using the method described by DeLong and colleagues [39].

Results

A Bayesian network structure was defined based on expert domain knowledge and can be found in Fig. 1A. The results of the 5-fold cross-validation can be found in Fig. 3. The AUC of the model was 0.61 (95% CI: 0.51–0.70).

Subsequently, we learned the model on all data, except for the data of a single hospital. The model is subsequently validated on the last hospital, yielding an ROC curve for every center. The result of this experiment can be observed in Fig. 4. The AUC of the model validated on the Maastro dataset is 0.59 (95% CI: 0.49–0.69). The AUC of the model validated on the Jessa dataset is 0.66 (95% CI: 0.36–0.95). The AUC of the model validated on the Liege dataset is 0.69 (95% CI: 0.53–0.84). The AUC of the model validated on the Aachen dataset is 0.63 (95% CI: 0.36–0.90). The AUC of the model validated on the Eindhoven dataset is 0.71 (95% CI: 0.41–1). The CPTs of the learned model can be seen in Table 2.

We compared the performance of models learned in a distributed fashion to models learned locally on the same data for either 1000, 10,000 or 10,000 patients partitioned over 10 to 100 subsets with 0, 20% and 40% randomly missing data. The results of this analysis can be observed in Fig. 5. The average difference of percentages in the CPTs between the distributed model and local model averaged over all levels of degradation and all numbers of hospitals is 1.6% (±0.7%) for dataset size of 100,000. The average difference in terms of AUC between the distributed model and local model averaged over all levels of degradation and all numbers of hospitals is 0.002 (±0.002) for dataset size of 100,000.

We compared the performance of the model for 3 different network structures. First, the expert defined network structure (Fig. 1A) was compared to the same structure without chemo timing. The AUC of the model is higher without chemo timing (AUC = 0.75, 95% CI: 0.53–0.97 versus 0.71 95% CI: 0.42–1), however, this difference is not significant ($P = 0.84$). We have used a data-driven approach to determine the network structure (Fig. 1B). The performance of the model using this structure is lower (AUC = 0.66, 95% CI: 0.36–0.96). However, this difference is not significant for the expert defined structure ($P = 0.8$), nor is it for the expert defined structure without chemo timing ($P = 0.37$).

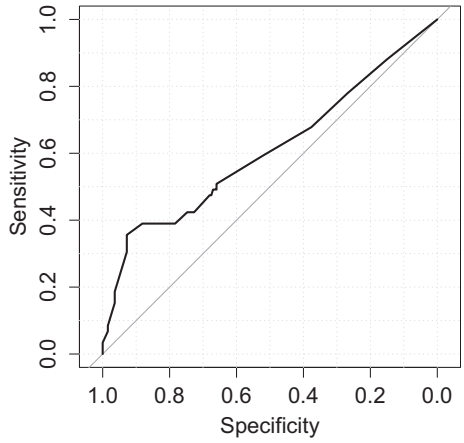

Fig. 3. Receiver operator characteristic on a 5-fold cross-validation. The Bayesian network uses chemo timing (no chemo, sequential or concurrent), tumor location (right lower lobe, right hilus, right upper lobe), forced expiratory volume in 1 s, in %, adjusted for age and gender; measured prior to medication, pre-treatment dyspnea, baseline dyspnea score (CTCAE grade < 2) and cardiac comorbidity (Non-hypertension cardiac disorder (at baseline)) to classify acute dyspnea.

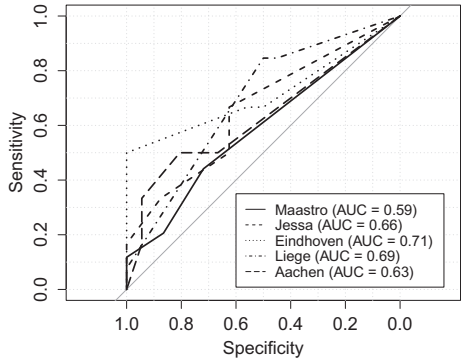

Fig. 4. Receiver operator characteristic curves for the model trained on all but one hospital and validated on the remaining hospital.

Discussion

In this study, we developed and implemented a distributed learning approach for Bayesian networks using data from 5

Please cite this article in press as: Jochems A 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), http://dx.doi.org/10.1016/j.radonc.2016.10.002

Table 2
Conditional probability tables of the final model. Associated structure can be observed in Fig. 1. An additional table is added for Dyspnea given the same structure as observed in Fig. 1A, without chemo timing.

FEV 1						
Location	Hilar Area of the Right Lung		Upper Lobe of the Right Lung		Lower Lobe of the Right Lung	
Cardiac comorbidity	No	Yes	No	Yes	No	Yes
Low (<30)	32%	27%	14%	21%	23%	8%
Medium (between 30 and 80)	8%	16%	10%	21%	19%	41%
High (>80)	61%	57%	76%	58%	58%	51%

Dyspnea						
Baseline dyspnea	<Grade 2			⩾Grade 2		
Chemo timing	None	Sequential	Concurrent	None	Sequential	Concurrent
<Grade 2	78%	61%	76%	25%	25%	16%
⩾Grade 2	22%	39%	24%	75%	75%	84%

Dyspnea (without chemo timing as parent)		
Baseline dyspnea	<Grade 2	⩾Grade 2
<Grade 2	72%	30%
⩾Grade 2	28%	70%

Location	
Hilar Area of the Right Lung	22%
Upper Lobe of the Right Lung	52%
Lower Lobe of the Right Lung	25%

Chemo timing	
None	13%
Sequential	18%
Concurrent	69%

Cardiac comorbidity	
No	72%
Yes	28%

Baseline dyspnea			
Location	Hilar Area of the Right Lung	Upper Lobe of the Right Lung	Lower Lobe of the Right Lung
<Grade 2	94%	95%	98%
⩾Grade 2	6%	5%	2%

hospitals in 3 countries and 3 languages without the need for any data to leave the individual hospital. We validated the performance of the model by cross-validation and by external validation in one hospital that was not included in the training. Finally, we compared the results of the distributed algorithm to the results of the centralized algorithm on the same data.

Implementation of the infrastructure

Distributed learning for the development of predictive models for radiotherapy is a relatively new topic, although a frameworks for international research data exchange have been proposed [40,41]. A project that has successfully made use of data sharing in a distributed manner is euroCAT (www.eurocat.info; watch the animation at https://youtu.be/ZDJFOxpwqEA), a collaborative project involving radiotherapy institutes from the Netherlands, Germany and Belgium. To our knowledge, no publications exist that use Bayesian networks for distributed learning for radiotherapy.

The distributed learning approach circumvents a number of human barriers issues but also raises the technical hurdle of setting up the systems that enable distributed learning. We have encountered a number of problems associated with setting up these systems. First, extraction of the data to be used for distributed learning still needs to be retrieved from the various medical record- and treatment planning systems used in the hospital. Therefore, we needed extensive communication with the data managers at every institute to find the data and make sure it is extracted properly. Different hospitals store their data in different formats and standards, leading to problems. Some of these problems can be solved by inferencing based on semantic web technology. For example, ECOG performance score and Karnofsky status can be mapped and made interoperable via the ontology, based on existing literature on how the conversion should be done [42]. Conversions between different units of the metric system can also be handled by the ontology in an automated fashion. Some issues, however, cannot be solved in an automated fashion. For example, some hospitals have their toxicity grades described in a free-form text field. These free-from text fields need to be read by human eyes and converted into a database before mapping can take place. This is the most time consuming aspect of installing the infrastructure. Once the data to be extracted is identified and located, semantic web technologies enable seamless mapping of local language terms to a universal language that the computer understands. The second issue we were confronted with was reluctance of IT staff to open the installed infrastructure for learning over the internet. The system administrators at the hospitals are cautious and at first unwilling to open up ports on the firewall for arbitrary reasons. A detailed explanation of the workings of the system was often required to convince them that there are no security risks involved with implementing the infrastructure.

Please cite this article in press as: Jochems A 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), http://dx.doi.org/10.1016/j.radonc.2016.10.002

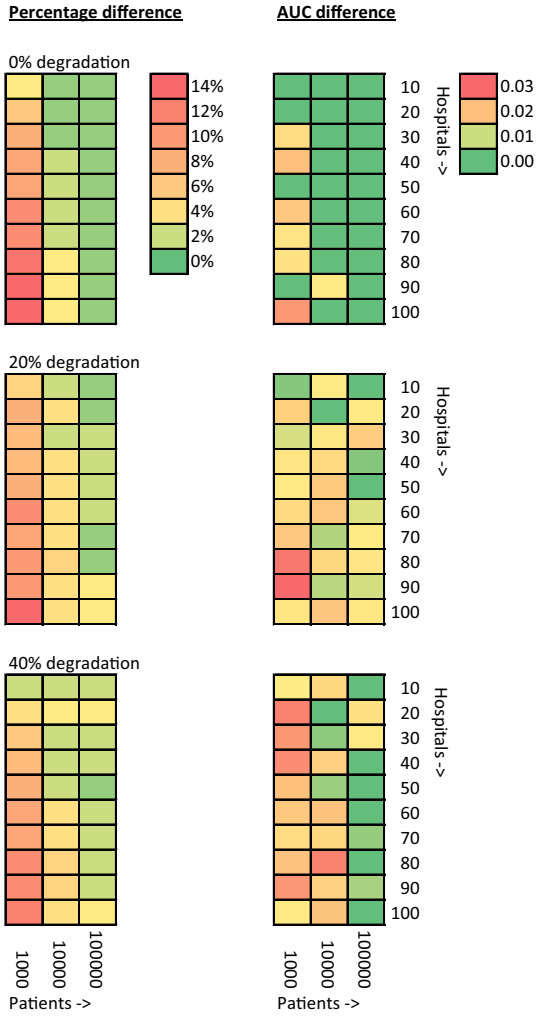

Fig. 5. Comparison of distributed learning versus local learning. Datasets were created by random sampling from the MAASTRO clinic data ($N = 123$). The first column shows the average difference in percentages of the conditional probability tables for the global and distributed model. The second column shows the difference in AUC for the global and distributed model. Rows depict the levels of artificially introduced random missing data (0%, 20%, 40%, respectively).

Distributed learning using regression and SVMs

Techniques for distributed privacy-preserving linear, log-linear and logistic regression have already been proposed. Our method has an advantage over the earlier proposed methods. Bayesian networks do not require separate imputation methods to handle missing data. They circumvent this problem by Bayesian statistics which is important as hospital datasets often contain numerous missing values as we demonstrated in previous work [43]. The capacity to handle missing values properly is therefore crucial. As the goal of distributed learning is to learn from large numbers of patient data, using a method that is relatively fast is essential. Our algorithm finished learning from the 268 patients in less than 5 min. The bottleneck of the system was transmission of models back and forth between the central location and the hospitals, which accounted for over 80% of the time.

Distributed learning in Bayesian networks

Numerous algorithms for distributed learning for Bayesian networks have been proposed, for both horizontally and vertically distributed data with various levels of privacy protection [11–14,30]. Most of these studies focus on learning the Bayesian network structure. Our work is an important addition to these existing studies, as our proposed method deals with the learning of the CPTs, rather than learning the network structure. Meng and colleagues propose a method by which distributed CPT learning can be done on horizontally distributed data [20]. However, their algorithm assumes all variables are binary valued. This is a restriction that is detrimental to model performance. Our approach enables learning from discretized variables of any number of levels and included determining the optimal bins of such discretization. Other CPT learning methods have been proposed [19,45], however, these algorithms focus on vertically partitioned data. Our algorithm provides a solution for horizontally partitioned data.

Although that it can be observed from Fig. 5 that the models learned from distributed data have near-equivalent performance in comparison to that of models learned locally, a number of issues could arise in the future using our variable discretization approach. For example, a site with a low volume of patients may yield bin boundaries that are not representative due to overfitting. Additionally, bin boundaries cannot be computed in a hospital if there are no events available for the outcome in the data of that hospital. A number of solutions exist for problems with variable discretization. First, the experimenter can reject the bin boundaries computed in hospitals if they seem clinically implausible. Bin boundaries computed at the remaining sites may be combined to yet come to an acceptable global solution. Second, the binning strategy suggested in this paper may be replaced by setting bin boundaries manually based on locally learned data or observations in the literature.

Privacy preservation is of vital importance when learning from medical data, as patient privacy and confidentiality of the physician are at risk [46]. In our approach, no data are transmitted between hospitals, other than the average bin information and model parameters. This information is an aggregate statistic and does not convey individual patient data, as long as there are multiple patients stored at each hospital. The hospitals communicate with the central location, but again only aggregate statistics relevant to the model are sent back and forth. If we consider the central location to be a trusted source, our algorithm is privacy preserving. Care must be taken that a sufficient number of patients are available at each site, so that none can be identified on the basis of the aggregate statistics.

Model comparison

Previously, our group has developed a model for radiation-induced lung toxicity based on single site data [47]. The previous model performed with an AUC of 0.67 on an external validation set. The model in this study performs better in some hospitals (Eindhoven, Liege) and worse in others (Maastro, Aachen, Jessa). Both studies were done on a limited number of patients (268 in this study versus 259 in the previous study [47]). This in turn results in large confidence intervals mentioned in the results section.

It can be observed from the CPTs in Table 2 that baseline dyspnea greatly modulates the post-RT dyspnea score. The model predicts an 84% chance of dyspnea in the event that a patient has $\geqslant$grade 2 baseline dyspnea and concurrent chemotherapy, whereas a 24% chance of $\geqslant$grade 2 post-RT dyspnea is predicted for patients with <2 grade baseline dyspnea. This is in line with previous work on dyspnea prediction [47]. Removing timing of chemotherapy from the network structure gives a slight boost in performance, however this is not significant. This could be due to the small data set size, as other work indicates that chemo timing does modulate dysphagia occurrence [47,48]. Using a data-driven method to determine the network structure reduces model perfor-

Please cite this article in press as: Jochems A 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), http://dx.doi.org/10.1016/j.radonc.2016.10.002

mance, although this difference is non-significant. This is in contrast to a study conducted by Sesen and colleagues that indicates that data-driven network structures outperform experts [49]. Our conflicting findings could be because we are predicting different outcomes, use different variables and use different datasets. As can be observed in Fig. 1B, data-driven networks do not take into account the temporal ordering of information. All arrows point away from dyspnea, although this is the outcome to predict. Such a network structure may be less likely to be adopted by physicians because it makes little intuitive sense. Further research is required to determine the most suitable network structure for a particular cancer and treatment outcome.

A modeling study done by Oberije and colleagues on dyspnea prediction used age, WHO-performance scale, FEV1, nicotine use, and mean lung dose as variables to predict acute dyspnea [50]. The performance of this model leaves room for improvement (AUC of 0.61 on external validation). FEV is the only common variable used this study and in the work done by Oberije and colleagues which seems in line with the importance of pre-therapy dyspnea which is correlated with FEV1. The difference in chosen variables could be due to the difference in modeling techniques used for both studies. Logistic regression analysis is unable to model non-linear interactions between variables whereas Bayesian networks can.

The use of historical data from routine clinical practice for decisions concerning new patients or to test new hypotheses is known as rapid learning [4,43]. Rapid learning brings numerous advantages, such as the large number of available patient data and the reduced selection bias present in the data in comparison to that of clinical trials [43]. Using the distributed learning method presented in this study, we make a large stride toward the implementation of the rapid learning healthcare practice. Some hurdles need to be resolved to make proper use of the distributed learning approach. First, getting access to- and locating the data stored at the individual hospitals requires substantial time investment. Second, one has to make sure the data is properly mapped onto the ontology. Third, the IT staff at each hospital has to be convinced that the learning infrastructure is safe before it can be connected to the internet.

Conclusion

In this work, we have shown that it is possible to develop a Bayesian network model to predict dyspnea after radiotherapy treatment on distributed data of lung cancer patients. As future work, we intend to use the distributed learning method described in this study to train models to predict multiple outcomes for a wide variety of cancers based on large volumes of data originating from multiple sources.

Conflict of interest statement

MAASTRO receives research funding (outside of this project) from Varian Medical Systems (VATE project).

Acknowledgements

Authors acknowledge financial support from the Interreg grant euroCAT, the Dutch technology Foundation STW (grant n° 10696 DuCAT & n° P14-19 Radiomics STRaTegy), which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs. Authors also acknowledge financial support from the EU 7th framework program (ARTFORCE – n° 257144, REQUITE – n° 601826), SME Phase 2 (EU proposal 673780 – RAIL), the European Program H2020-2015-17 (BD2Decide – PHC30-689715 and ImmunoSABR – n° 733008), EUROSTARS (SeDI, CloudAtlas, DART), Kankeronderzoekfonds Limburg from the Health Foundation Limburg, Alpe d'HuZes-KWF (DESIGN) and the Dutch Cancer Society. We would like to thank Varian for providing the distributed learning manager and Wolfgang Wiessler for his dedicated support.

References

[1] Etheredge LM. A rapid-learning health system. Health Aff Proj Hope 2007;26: w107–18. doi: http://dx.doi.org/10.1377/hlthaff.26.2.w107.
[2] Lambin P, Petit SF, Aerts HJWL, van Elmpt WJC, Oberije CJG, Starmans MHW, et al. From population to voxel-based radiotherapy: exploiting intra-tumour and intra-organ heterogeneity for advanced treatment of non-small cell lung cancer. Radiother Oncol J Eur Soc Ther Radiol Oncol 2010;96:145–52. doi: http://dx.doi.org/10.1016/j.radonc.2010.07.001.
[3] Lambin P, Zindler J, Vanneste B, Van De Voorde L, Jacobs M, Eekers D, et al. Modern clinical research: How rapid learning health care and cohort multiple randomised clinical trials complement traditional evidence based medicine. Acta Oncol 2015;54:1289–300.
[4] Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, et al. Rapid-learning system for cancer care. J Clin Oncol 2010;28:4268–74. doi: http://dx.doi.org/10.1200/JCO.2010.28.5478.
[5] Lambin P, Zindler J, Vanneste BGL, De Voorde LV, Eekers D, Compter I, et al. Decision support systems for personalized and participative radiation oncology. Adv Drug Deliv Rev 2016. doi: http://dx.doi.org/10.1016/j.addr.2016.01.006.
[6] Jayasurya K, Fung G, Yu S, Dehing-Oberije C, De Ruysscher D, Hope A, et al. Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. Med Phys 2010;37:1401–7.
[7] Dehing-Oberije C, De Ruysscher D, Petit S, Van Meerbeeck J, Vandecasteele K, De Neve W, et al. Development, external validation and clinical usefulness of a practical prediction model for radiation-induced dysphagia in lung cancer patients. Radiother Oncol 2010;97:455–61.
[8] Lambin P, van Stiphout RGPM, Starmans MHW, Rios-Velazquez E, Nalbantov G, Aerts HJWL, et al. Predicting outcomes in radiation oncology–multifactorial decision support systems. Nat Rev Clin Oncol 2013;10:27–40. doi: http://dx.doi.org/10.1038/nrclinonc.2012.196.
[9] Doshi P, Jefferson T, Del Mar C. The imperative to share clinical study reports: recommendations from the Tamiflu experience. PLoS Med 2012;9:e1001201. doi: http://dx.doi.org/10.1371/journal.pmed.1001201.
[10] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. Ann Intern Med 2015;162:55–63. doi: http://dx.doi.org/10.7326/M14-0697.
[11] Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends® Mach Learn 2011;3:1–122.
[12] Karr AF, Lin X, Sanil AP, Reiter JP. Secure regression on distributed databases. J Comput Graph Stat 2005;14:263–79. doi: http://dx.doi.org/10.1198/106186005X47714.
[13] Karr AF, Lin X, Sanil AP, Reiter JP. Privacy-preserving analysis of vertically partitioned data using secure matrix products. J Off Stat 2009;25:125.
[14] Sanil AP, Karr AF, Lin X, Reiter JP. Privacy preserving regression modelling via distributed computation. In: Proc Tenth ACM SIGKDD Int Conf Knowl Discov Data Min. ACM; 2004. p. 677–82.
[15] Chen R, Sivakumar K, Kargupta H. Learning Bayesian network structure from distributed data. In: Barbara D, Kamath C, editors. Proc 2003 SIAM Int Conf Data Min. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2003. p. 284–8.
[16] Na Y, Yang J. Distributed Bayesian network structure learning. Ind Electron ISIE 2010 IEEE Int Symp On IEEE, 2010. p. 1607–11.
[17] Gou KX, Jun GX, Learning Zhao Z. Bayesian network structure from distributed homogeneous data. Softw Eng Artif Intell Netw Parallel Distributed Comput 2007 SNPD 2007 Eighth ACIS Int Conf On IEEE 2007;3:250–4.
[18] Wright R, Yang Z. Privacy-preserving Bayesian network structure computation on distributed heterogeneous data. In: Proc Tenth ACM SIGKDD Int Conf Knowl Discov Data Min. New York, NY, USA: ACM; 2004. p. 713–8. doi: http://dx.doi.org/10.1145/1014052.1014145.
[19] Yang Z, Wright RN. Improved privacy-preserving bayesian network parameter learning on vertically partitioned data. 21st Int Conf Data Eng Workshop 2005 2005:1196. doi: http://dx.doi.org/10.1109/ICDE.2005.230.
[20] Meng D, Sivakumar K, Privacy-sensitive Kargupta H. Bayesian network parameter learning. Data Min 2004 ICDM04 Fourth IEEE Int Conf On IEEE 2004:487–90.
[21] Ruysscher DD, Dehing C, Yu S, Wanders R, Öllers M, Dingemans A-MC, et al. Dyspnea evolution after high-dose radiotherapy in patients with non-small cell lung cancer. Radiother Oncol 2009;91:353–9. doi: http://dx.doi.org/10.1016/j.radonc.2008.10.006.
[22] Jain S, Poon I, Soliman H, Keller B, Kim A, Lochray F, et al. Lung stereotactic body radiation therapy (SBRT) delivered over 4 or 11 days: a comparison of acute toxicity and quality of life. Radiother Oncol 2013;108:320–5. doi: http://dx.doi.org/10.1016/j.radonc.2013.06.045.

[23] Rodrigues G, Lock M, D'Souza D, Yu E, Van Dyk J. Prediction of radiation pneumonitis by dose–volume histogram parameters in lung cancer–a systematic review. Radiother Oncol J Eur Soc Ther Radiol Oncol 2004;71:127–38. doi: http://dx.doi.org/10.1016/j.radonc.2004.02.015.
[24] Oberije C, Liao Z, De Ruysscher D, Tucker S, Lambin P. Development and external validation of a model for prediction of radiation-induced dyspnea: an approach combining clinical data with information from literature. Int J Radiat Oncol Biol Phys 2010;78:S528.
[25] Allemang D, Hendler J. Semantic web for the working ontologist: effective modeling in RDFS and OWL. Elsevier; 2011.
[26] Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu W-L, Wright LW. NCI thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform 2007;40:30–43.
[27] Heath T, Bizer C. Linked data: evolving the web into a global data space. Morgen & Claypool Publishers; 2011.
[28] Rudolph S, Gottlob G, Horrocks I, van Harmelen F. Reasoning web. Semantic Technologies for intelligent data access: 9th International Summer School 2013. In: Mannheim, Germany, July 30–August 2, 2013. Proceedings. Springer; 2013.
[29] Casters M, Bouman R, Van Dongen J. Pentaho Kettle Solutions: building open source ETL solutions with Pentaho Data Integration. John Wiley & Sons; 2010.
[30] Bizer C, Seaborne A. D2RQ-treating non-RDF databases as virtual RDF graphs. In: Proc 3rd Int Semantic Web Conf ISWC2004, Vol. 2004. Citeseer Hiroshima; 2004.
[31] Broekstra J, Kampman A, van Harmelen F. Sesame: a generic architecture for storing and querying RDF and RDF schema. In: Horrocks I, Hendler J, editors. Semantic web — ISWC 2002. Berlin Heidelberg: Springer; 2002. p. 54–68.
[32] Quilitz B, Leser U. Querying distributed RDF data sources with SPARQL. Springer; 2008.
[33] Lauritzen SL. The EM algorithm for graphical association models with missing data. Comput Stat Data Anal 1995;19:191–201.
[34] Kuschner KW, Malyarenko DI, Cooke WE, Cazares LH, Semmes OJ, Tracy ER. A Bayesian network approach to feature selection in mass spectrometry data. BMC Bioinformatics 2010;11:177.
[35] Spirtes P, Glymour C. An algorithm for fast recovery of sparse causal graphs. Soc Sci Comput Rev 1991;9:62–72.
[36] Druzdzel MJ. SMILE: structural modeling, inference, and learning engine and GeNIe: a development environment for graphical decision-theoretic models. AAAI/IAAI 1999:902–3.
[37] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. PROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77.
[38] LeDell E, Petersen ML, van der Laan MJ. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates, 2012.
[39] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.
[40] Roelofs E, Dekker A, Meldolesi E, van Stiphout RGPM, Valentini V, Lambin P. International data-sharing for radiotherapy research: an open-source based infrastructure for multicentric clinical data mining. Radiother Oncol 2014;110:370–4. doi: http://dx.doi.org/10.1016/j.radonc.2013.11.001.
[41] Skripcak T, Belka C, Bosch W, Brink C, Brunner T, Budach V, et al. Creating a data exchange strategy for radiotherapy research: towards federated databases and anonymised public datasets. Radiother Oncol 2014;113:303–9. doi: http://dx.doi.org/10.1016/j.radonc.2014.10.001.
[42] Oken MM, Creech RH, Tormey DC, Horton J, Davis TE, McFadden ET, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. Am J Clin Oncol 1982;5:649–55.
[43] Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CML, et al. Rapid Learning health care in oncology – an approach towards decision support systems enabling customised radiotherapy. Radiother Oncol 2013;109:159–64. doi: http://dx.doi.org/10.1016/j.radonc.2013.07.007.
[45] Ma J, Sivakumar K. Privacy-preserving Bayesian network learning from heterogeneous distributed data. DMIN. Citeseer; 2006. p. 246–52.
[46] El Emam K, Hu J, Mercer J, Peyton L, Kantarcioglu M, Malin B, et al. A secure protocol for protecting the identity of providers when disclosing data for disease surveillance. J Am Med Inform Assoc JAMIA 2011;18:212–7. doi: http://dx.doi.org/10.1136/amiajnl-2011-000100.
[47] Nalbantov G, Kietselaer B, Vandecasteele K, Oberije C, Berbee M, Troost E, et al. Cardiac comorbidity is an independent risk factor for radiation-induced lung toxicity in lung cancer patients. Radiother Oncol 2013;109:100–6.
[48] Dehing-Oberije C, De Ruysscher D, van Baardwijk A, Yu S, Rao B, Lambin P. The importance of patient characteristics for the prediction of radiation-induced lung toxicity. Radiother Oncol J Eur Soc Ther Radiol Oncol 2009;91:421–6. doi: http://dx.doi.org/10.1016/j.radonc.2008.12.002.
[49] Sesen MB, Nicholson AE, Banares-Alcantara R, Kadir T, Brady M. Bayesian networks for clinical decision support in lung cancer care. PLoS ONE 2013;8: e82349. doi: http://dx.doi.org/10.1371/journal.pone.0082349.
[50] Oberije C, Liao Z, De-Ruysscher D, Tucker S, Lambin P. Development and external validation of a model for prediction of radiation-induced dyspnea: an approach combining clinical data with information from literature. Int J Radiat Oncol Biol Phys 2010;78:S528. doi: http://dx.doi.org/10.1016/j.ijrobp.2010.07.1233.

Please cite this article in press as: Jochems A 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), http://dx.doi.org/10.1016/j.radonc.2016.10.002