Implementation of the Australian Computer-Assisted Theragnostics (AusCAT) network for radiation oncology data extraction, reporting and distributed learning

RADIATION ONCOLOGY—ORIGINAL ARTICLE

Implementation of the Australian Computer-Assisted Theragnostics (AusCAT) network for radiation oncology data extraction, reporting and distributed learning

Matthew Field,[1,2] Shalini Vinod,[1,2,3] Noel Aherne,[4,5] Martin Carolan,[6] Andre Dekker,[7] Geoff Delaney,[1,2,3] Stuart Greenham,[4] Eric Hau,[8,9] Joerg Lehmann,[10,11,12] Joanna Ludbrook,[11] Andrew Miller,[6] Angela Rezo,[13] Jothybasu Selvaraj,[1,13] Jonathan Sykes,[8,12] Lois Holloway[1,2,3,12] and David Thwaites[12]

1 South Western Sydney Clinical School, Faculty of Medicine, UNSW, Sydney, New South Wales, Australia
2 Ingham Institute for Applied Medical Research, Liverpool, New South Wales, Australia
3 Liverpool and Macarthur Cancer Therapy Centres, Liverpool, New South Wales, Australia
4 Mid North Coast Cancer Institute, Coffs Harbour, New South Wales, Australia
5 Rural Clinical School, Faculty of Medicine, University of New South Wales, Sydney, New South Wales, Australia
6 Illawarra Cancer Care Centre, Wollongong, New South Wales, Australia
7 Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University, Maastricht, The Netherlands
8 Sydney West Radiation Oncology Network, Sydney, Australia
9 Westmead Clinical School, University of Sydney, Sydney, New South Wales, Australia
10 School of Mathematical and Physical Sciences, University of Newcastle, Newcastle, New South Wales, Australia
11 Department of Radiation Oncology, Calvary Mater, Newcastle, New South Wales, Australia
12 Institute of Medical Physics, School of Physics, University of Sydney, Sydney, New South Wales, Australia
13 Canberra Health Services, Canberra, Australian Capital Territory, Australia

M Field PhD; **S Vinod** MBBS MD FRANZCR; **N Aherne** MB BCh AFRCSI FRANZCR; **M Carolan** PhD; **A Dekker** PhD; **G Delaney** MBBS MD PhD FRANZCR; **S Greenham** GradDip(IT) BAppSci(Comp) DipMgt DipAppSci; **E Hau** BSc(Med) MBBS FRANZCR PhD; **J Ludbrook** MBChB FRANZCR; **A Miller** BMed BSc GradDIpEdMInfCommTech (Res) FRANZCR FAIDH; **A Rezo** MBBS BSc (Med) FRANZCR; **J Selvaraj** PhD; **J Sykes** PhD; **L Holloway** PhD; **D Thwaites** MA MSc PhD FInstP FIPEM FRCR(hon) FACPSEM.

Correspondence
Matthew Field, Ingham Institute for Applied Medical Research, 1 Campbell Street, Liverpool, NSW, 2170, Australia.
Email: matthew.field@unsw.edu.au

Conflict of interest: E Hau is on advisory panel and receives Honoria and research grant funding from Astra Zeneca. A Dekker works with Varian, Philips and Medical Data Works on similar network initiatives.

Summary

Introduction: There is significant potential to analyse and model routinely collected data for radiotherapy patients to provide evidence to support clinical decisions, particularly where clinical trials evidence is limited or non-existent. However, in practice there are administrative, ethical, technical, logistical and legislative barriers to having coordinated data analysis platforms across radiation oncology centres.

Methods: A distributed learning network of computer systems is presented, with software tools to extract and report on oncology data and to enable statistical model development. A distributed or federated learning approach keeps data in the local centre, but models are developed from the entire cohort.

Results: The feasibility of this approach is demonstrated across six Australian oncology centres, using routinely collected lung cancer data from oncology information systems. The infrastructure was used to validate and develop machine learning for model-based clinical decision support and for one centre to assess patient eligibility criteria for two major lung cancer radiotherapy clinical trials (RTOG-9410, RTOG-0617). External validation of a 2-year overall survival model for non–small cell lung cancer (NSCLC) gave an AUC of 0.65 and C-index of 0.62 across the network. For one centre, 65% of Stage III NSCLC patients did not meet eligibility criteria for either of the two practice-changing clinical trials, and these patients had poorer survival than eligible patients (10.6 m vs. 15.8 m, $P = 0.024$).

Conclusion: Population-based studies on routine data are possible using a distributed learning approach. This has the potential for decision support models for patients for whom supporting clinical trial evidence is not applicable.

Journal of Medical Imaging and Radiation Oncology

Lois Holloway and David Thwaites are joint senior authors.

Submitted 9 March 2021; accepted 29 June 2021.

doi:10.1111/1754-9485.13287

Key words: artificial intelligence; decision support systems; distributed learning; federated learning; radiation oncology.

Introduction

Randomised clinical trials (RCTs) provide level I evidence for efficacy of medical treatments, leading to clinical practice guidelines. However, trials have strict eligibility criteria, to detect differences with as few confounding factors as possible. This leaves a potential evidence gap for a relatively large group of patients who might not meet those criteria. Approximately 3% of patients are enrolled in cancer clinical trials.[1] Knowing the proportion of patients in the wider population who meet specific trial eligibility criteria would indicate the number for whom trial-based guidelines are directly applicable. However, this is generally unknown, nor it is known whether they are treated differently from trial-based recommendations.[2] 'Closing the translation loop' is also important to assess the impact of RCT-derived guidelines on 'real-world' clinical practice, including for patients not meeting RCT eligibility criteria where clinicians have to extrapolate from RCTs and choose whether or not to apply trial-based evidence for such patients.

Using routine clinical data collected in electronic health records (EHRs) provides a means of analysing all treated patients,[3–5] including determining how many meet given trial eligibility requirements. This enables comparison of, and learning from, different treatment approaches and assessment of translation of trial evidence into clinical practice. Figure 1 provides a schematic description of the potential data available, considering clinical trials and clinical practice data and two patient scenarios, where clinical trial eligibility criteria apply or not.

Many cancer centre research groups have combined efforts to use EHR data to learn from population-level outcomes and to develop clinical decision support systems. Such systems based on population-based learning will be more applicable than those based solely on RCT data.[6,7] However, there are multiple logistical, ethical, administrative, legal and technical hurdles to achieving consistent, reproducible analyses of population-based radiation oncology data, especially across jurisdictions.

The Australian Computer Assisted Theragnostics network (AusCAT, formerly OzCAT) was established, in initial collaboration with the pioneering MAASTRO group, to overcome some of these hurdles. Development began with some single-centre studies and gradually expanded to include more centres as outlined below.

In preliminary work, Dekker et al.[8] validated a non–small cell lung cancer (NSCLC) two-year overall survival model[9] on a single-centre Australian cohort. Lustberg et al.[10] repeated this for laryngeal carcinoma at a different Australian centre and compared results to an RTOG trial dataset. These studies demonstrated the feasibility of automated data extraction of routine data from the clinical systems in those centres and assessed missing data proportions for the considered model parameters. For the NSCLC study, data were missing for 62% of

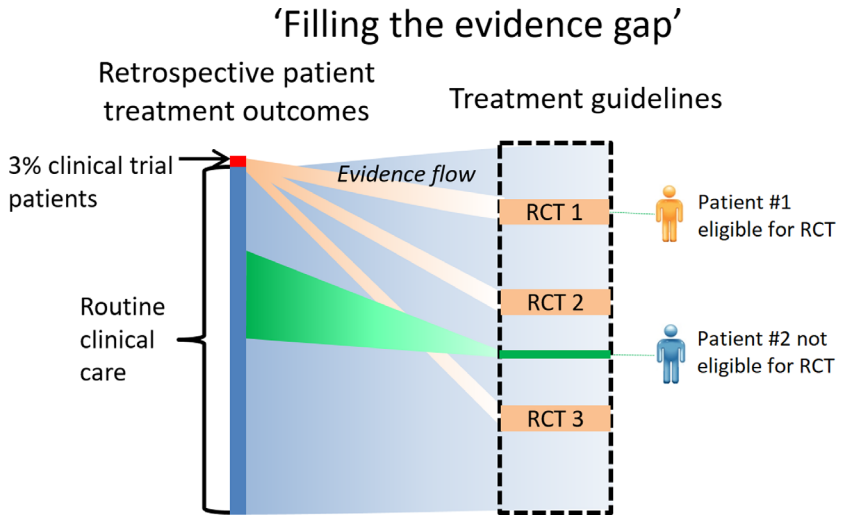

Fig. 1. Schematic of data and evidence available to support treatment decisions. Patient #1 is directly supported by clinical trial evidence. Patient #2 is provided personalised decision support, utilising relevant retrospective data from routine clinical care.

patients for some key parameters (e.g. gross tumour volume (GTV) and 2-year overall survival) for this 1995–2013 cohort. However valid prediction models were still achievable. The laryngeal carcinoma study demonstrated that validating the model on clinical practice data, even with missing data, exhibited improved risk stratification of patients compared to the RTOG trial dataset.

Australian Computer-Assisted Theragnostics has subsequently been extended from single to multiple centres by utilising distributed learning, where patient datasets remain at the local institution and models are generated by exchanging model parameters (not data) from each centre. This concept has been demonstrated in various software platforms and collaborations, including the Varian Learning Portal (VLP),[11,12] Personal Health Train infrastructure (PHT),[13] DataSHIELD,[14] and WebDISCO.[15] The current work presents the first implementation of this concept across Australian centres, using a software platform developed in-house to provide maximum flexibility and clear local management and control of its development and modification. Work began on this in 2014.[16] A NSCLC cohort receiving radiotherapy (RT) is selected as the demonstration example and the presented results are chosen to illustrate the overall function and potential of distributed machine learning.

The network's initial aims were to demonstrate the feasibility of: automatically extracting, de-identifying and standardizing datasets, including imaging, from clinical systems across the centres; assessing data availability and quality for this patient cohort; securely and efficiently developing and validating machine learning–based outcome-prediction models. In this article, an overall survival model is externally validated for patients with unresectable Stage I–III NSCLC treated with radiotherapy.[9] For one centre as an example, an additional aim was to identify the proportion of Stage III NSCLC patients who did not meet the eligibility criteria for the main lung cancer RT RCTs on which the NSW Cancer Institute EviQ guidelines[17] are based and to assess any difference in outcome for these patients. Various lung cancer prognostication models have been published[9,18,19]; however, we are unaware of previous work to assess the potential patient population not meeting the RCT criteria on which treatment guidelines are based.

Methods

The AusCAT network was established across six radiation oncology treatment centres in New South Wales (NSW) and Australian Capital Territory (ACT). This comprises facilities in eleven hospitals, since some centres are multisite. They are Illawarra Cancer Care Centre, Shoalhaven Cancer Care Centre, Crown Princess Mary Cancer Centre (Westmead Hospital), Blacktown Hospital Cancer Centre, Liverpool Cancer Therapy Centre, Macarthur/Campbelltown Therapy Centre, Mid North Coast Cancer Institute (Port Macquarie, Coffs Harbour), North Coast Cancer Institute (Lismore), Canberra Hospital, and Newcastle Calvary Mater Hospital. Project approval was granted by the NSW Population and Health Services Research Ethics Committee (HREC/16/CIPHS/5).

Patient data

For each centre a local research database was established to house the extracted data.[16]

The patient cohort selected for initial feasibility evaluation and for validating and developing decision support models was for lung cancer diagnosis according to ICD-10 and ICD-9 codes and treated with RT. Selected patients were Stage I–IIIB NSCLC, with no prior surgery or prior thoracic RT and excluding those treated with stereotactic body radiation therapy (SBRT). Total radiotherapy dose was used as a decision variable, defining treatment with >45 Gy as curative intent RT and below this threshold, palliative intent.[9]

For the model considered, the required variables included forced expiratory volume (FEV1%), age, gender, performance status, number of positive lymph node stations and GTV. Where data were not available, the recommended Bayesian network imputation approach for the published model[20] was used. Additional variables were needed for the RCT eligibility analysis, including weight loss, chemotherapy, pathology results (e.g. serum creatinine, bilirubin, platelet count, hemoglobin, absolute neutrophil count), respiratory and cardiovascular comorbidities, neuropathy, pleural effusion and histologic confirmation.

Data were sourced from each cancer centre's oncology information systems (OIS) database, either Mosaiq (Elekta) or Aria (Varian), using structured query language (SQL) queries and preparing data into a reportable format stored in an anonymized database at each clinic.[16] Customized text-mining with a set of regular expressions was used to extract further information from free-text clinical notes to supplement the structured data for performance status, FEV1%, smoking status, and weight loss. The combined GTV was used, including primary and nodal regions where available. For this, automated DICOM-RT exports were facilitated for CT and RTSTRUCT from the treatment planning archives (XiO [Elekta], Monaco [Elekta], Pinnacle [Philips], and Eclipse [Varian]). The DICOM information was anonymised with the RSNA Clinical Trial Processor (CTP) and stored locally in a picture archiving and communications system (PACS). In-house software retrieved the DICOM files and selected or combined the necessary regions to compute the combined GTV based on a curated list of approved structure names.

For four of the six centres, the data were supplemented by additional information from the Registry of Births, Deaths and Marriages to update 2-year overall

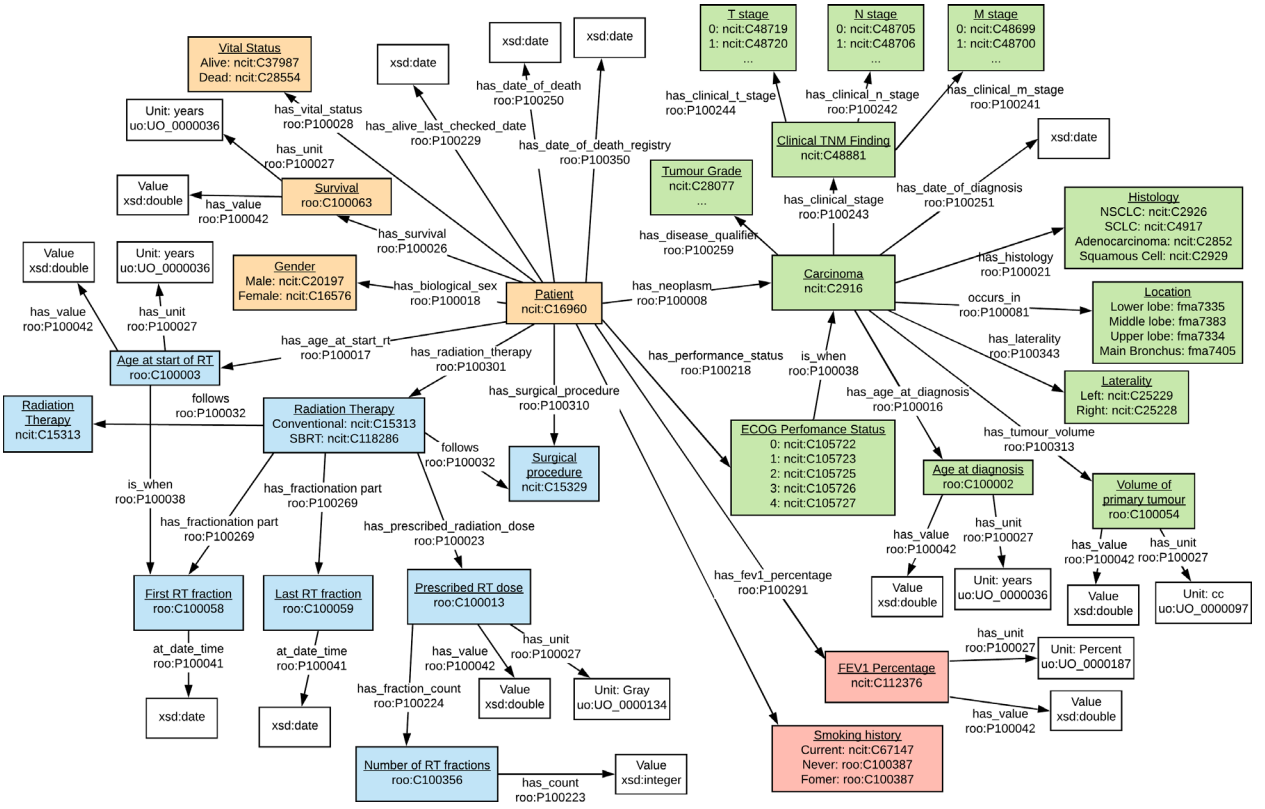

Fig. 2. Structure of the NSCLC data set definitions linking to NCI thesaurus and ROO. From the centre of the graph the patient and demographic data are highlighted in orange, disease related data are highlighted in green, treatment related data are highlighted in blue and additional patient measurements are in red. Not all variables are shown due to space constraints.

survival data and Admitted Patient Data Collection to update surgical events prior to RT, linking data via the Centre for Health Record Linkage (CHeReL).

The data were translated into a further standardized format using available data dictionaries (ontologies) from the National Cancer Institute[21] and others, e.g. the radiation oncology ontology (ROO)[22] to create a findable, accessible, interoperable and reusable (FAIR) data model.[23,24] Minor additions were made to the dictionary to describe the retrieved data for this project. Figure 2 presents a graphic visualization of the FAIR data model.

Distributed learning software

A custom software platform for distributed learning was developed for AusCAT network use.[16] It consists of a set of Java web services to coordinate the communication of algorithms, models, and statistics aggregated over patients, between the clinic systems and a central server located at Ingham Institute for Applied Medical Research. The code was developed in MATLAB and was designed to send a compiled algorithm to each clinic storing the datasets and allow the algorithm to then send and receive locally derived model parameters and statistics. Some of the analyses required statistical approximations to be shared on the network, but only as aggregated information over groups of patients. Thus, each centre securely retains its own data while contributing to the overall model.

Survival prediction model validation

A support vector machine model predicting 2-year overall survival of Stage I–IIIB NSCLC patients treated with curative RT[9] was externally validated. This was performed on the overall cohort and the curative and palliative cohorts separately, computing the AUC for predicting 2-year overall survival and the concordance index (C-index). Kaplan–Meier plots were generated for the risk groups defined by the model, partitioned as low, medium, and high risk in the study by Dehing-Oberije C et al.,[9] and compared to predictions based on stage. The model work[9] combined the two middle quartiles of the patient distribution into the medium risk group; whereas here it was split back into quartiles named medium-low and medium-high at the midpoint between these two quartile thresholds. To assess treatment decision- making potential the risk groups were dichotomized into 'good prognosis' and 'poor prognosis'. For this, low and medium-low risk, or Stage I–II were

defined as good prognosis, with the other groups in each approach (model or staging) considered poor prognosis. The survival curve statistics were extracted by the distributed learning network in fixed two-week time windows over a 7-year period, where the two-week binning was to aggregate survival information across the cohort to ensure that we are not sharing individual data points across the network. To assess survival difference significance for each pair of risk curves, the log-rank test was applied with statistical significance defined as $P < 0.05$.

RCT eligibility

To further demonstrate data extraction applications, at one cancer centre patients were evaluated against eligibility criteria for two clinical trials; RTOG-9410 which confirmed the superiority of concurrent over sequential chemotherapy and RT for Stage III NSCLC, and RTOG-0617 which confirmed 60 Gy in 30 fractions as the standard radiotherapy dose for Stage III NSCLC. These form the basis of the EviQ NSCLC RT conventional fractionation guidelines.[25,26] The proportion of routine practice RT patients not meeting these RCTs' eligibility criteria was examined by retrieving the required clinical data from the OIS. Due to incomplete data in individual patient files, it was more straightforward to consider the numbers of patients not eligible for either RCT, since any single data item present that would exclude a patient from the trials indicates this directly and this approach gives a lower bound on the true proportion when complete data exists.

The proportion of patients receiving curative RT, their median survival and the survival curves were compared between those not meeting eligibility requirements for either trial and those potentially eligible for either trial. Due to the range of missing data and potential for changes in treatment practice influencing baseline survival rate, the cohort timepoints were varied to check the consistency of the log-rank test.

Results

Data retrieval

The total extracted patient cohort is 12 047 patients from the six centres. Figure 3a displays the distribution of patients over the network from 1996–2018, showing some variation in data availability among centres. Clinic A has data consistently stored in OIS (and thus in an extractable format) over the longest time period versus Clinic E, for example, where datasets are available only from 2011. In Figure 3c,d the rate of curative and palliative treatments across the network is displayed from the same period. The trend towards more aggressive treatment with time is driven by the majority of centres, although Figure 3d is given as an example of a centre that varies from the general group behaviour.

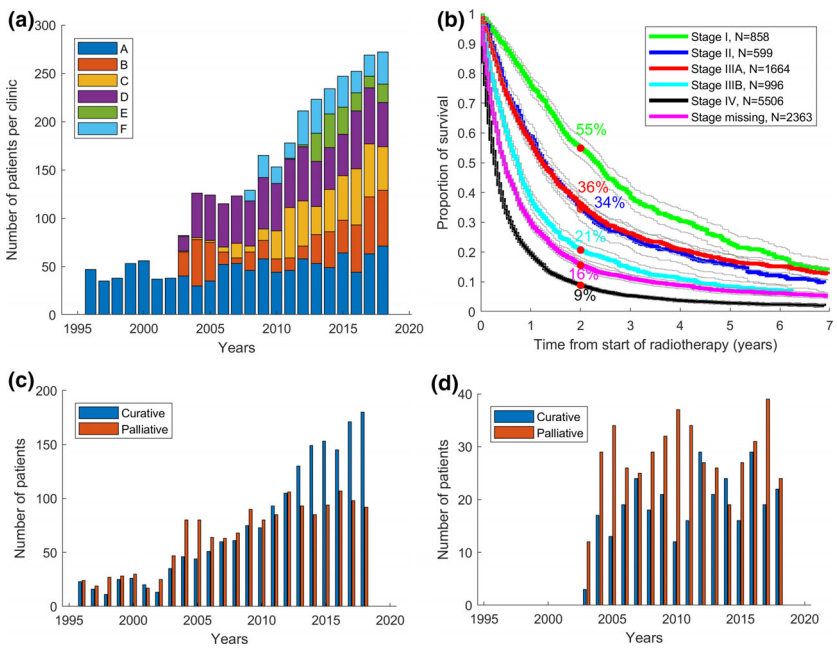

Fig. 3. (a) Lung cancer patients treated with RT across 6 radiation oncology centres (A–F). (b) Survival curves of the overall lung cohort stratified by stage including missing values for stage (c) NSCLC Stage IIIIB patients treated with RT across 6 centres per year grouped by treatment intent where curative intent is >45 Gy, noting that patients receiving SABR are excluded from this cohort. In (d) the same data is shown for Centre D, as an example of variation from the overall group behaviour, since this centre has less aggressive treatment and the palliative and curative rates track each other.

Table 1. Data availability in terms of meeting the model validation criteria and having all relevant data retrievable from databases. Also, comparison of model validation metrics for stage group prediction and for model prediction (AUC at two years and C-index for each centre and overall, where AUC is calculated over all centres and C-index is the mean), indicating superior performance for model prediction

Centre	A	B	C	D	E	F	Total
Lung-cancer RT Patients – n	3650	1986	1888	2893	597	1033	12047
Patients who met study selection criteria - *n* (%)	1100 (30)	477 (24)	448 (24)	754 (26)	146 (24)	272 (26)	3207 (27)
Missing stage or histology – n (%)	560 (15)	708 (36)	726 (38)	615 (21)	242 (41)	101 (10)	2952 (25)
Validation cohort curative RT 2013–2018 – n	250	184	179	131	78	106	928

							Combined/mean
Stage group AUC at two years	0.52	0.57	0.53	0.60	0.45	0.58	0.54
Stage group C-index	0.52	0.54	0.54	0.53	0.51	0.58	0.56
Model validation AUC at two years	0.65	0.64	0.66	0.71	0.61	0.63	0.66
Model validation C-index	0.58	0.59	0.56	0.63	0.76	0.58	0.62

In Table 1, the proportions of patients meeting the criteria for validation of the survival prediction model per clinic and in total are described. Of the 12 047 lung cancer patients treated with RT, 3207 (27%) were included and 2952 (25%) were excluded because of missing stage or histology data; the rest were mainly Stage IV or not NSCLC.

Model validation and decision support

An overall survival prediction model was externally validated and compared with using overall stage grouping for prediction, for patients treated between 2013 and 2018. Table 1 shows the AUC metric for each assessed centre at the 2-year survival endpoint and the C-index.

In Figure 4, the Kaplan–Meier curves for each overall approach (model and staging-based) are dichotomized into the risk groups defined by each approach in (a)–(b) and then as their predicted good and poor prognosis groups and by actual treatment (curative, palliative) in (c)–(d). Survival differences were examined to determine if the model can highlight patient subgroups that may have benefited from decision support. For each approach, patients that were considered to have good prognosis yet received palliative intent treatment had increased survival with respect to patients in the predicted poor prognosis group ($P < 0.001$). For the Dehing-Oberije et al.[9] model there was a significant difference between predicted good prognosis and predicted poor prognosis patients receiving curative RT ($P < 0.001$), in contrast to using overall stage for prediction, where the log-rank *P*-value was 0.12.

RCT eligibility proportion

For Centre A, the AusCAT database was queried for Stage III NSCLC patients, receiving RT between 2007 and 2019, who did not undergo surgery or any previous thoracic RT or have SBRT. Of the 498 patients identified, 349 (70%) did not meet the RTOG-9410 eligibility criteria and 445 (89%) did not meet the RTOG-0617 criteria. A total of 327 (66%) did not meet the requirements for either trial.

Figure 5 shows the data availability for meeting trial eligibility criteria at a patient-by-patient level, to demonstrate the quantity of missing data.

Overall, 61% of the 498 patients received curative RT, 57% for patients ineligible for either trial and 70% for patients potentially eligible. The median survival of the ineligible cohort that were curatively treated was 16.7 months, compared to 18.9 months for the potentially eligible patients. Figure 6a presents the Kaplan–Meier survival curves for curatively treated patients that did and did not meet trial eligibility criteria. Under the log-rank test the survival of each cohort is significantly different ($P = 0.012$), with similar survival rates in the short term (<2 years), but inferior survival for non-eligible patients in the longer term (>3 years), reflecting other factors, such as overall condition of the patients in each cohort.

The *P*-values were assessed for varying cohort selection periods from 2000–2019 to 2013–2019 and are shown in Figure 6b. The log-rank test indicates non-statistical significance beyond 2011.

Discussion

Within the AusCAT network, 27% of lung cancer patients had the required data elements to enable validation of the NSCLC model considered here, whilst for 25% of patients this could not be assessed due to missing data. Although not ideal, particularly compared to RCT which achieve close to 100% of required data items, this dataset was adequate to externally validate a published survival model and has demonstrated improved patient risk

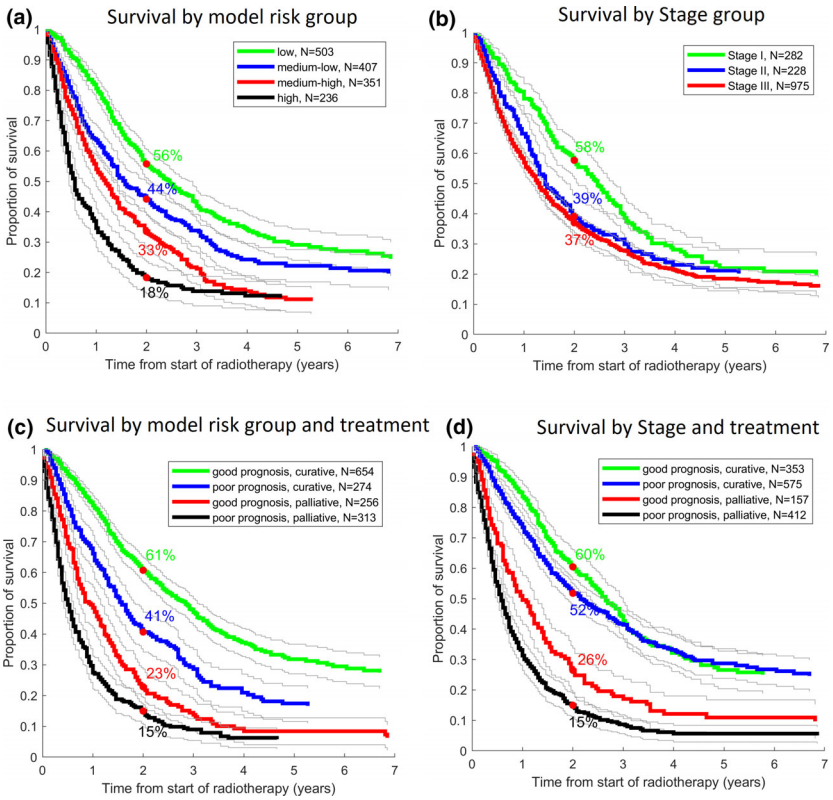

Fig. 4. Kaplan Meier survival curves stratified by risk group (a) as predicted by the externally validated model; and (b) by overall stage group. Survival curves are compared stratifying for good and poor prognosis risk groups and for treatment intent, (c) for the externally validated model; and (d) for overall stage. As an example of the terminology in the curve labelling key, in (c), 'good prognosis, palliative' means patients predicted by the model to have a good prognosis if treated radically, but who were actually treated palliatively.

stratification when compared to overall stage information alone.

Patients classified as good prognosis by the model, but who had received palliative treatment indicate that similar future cases could be considered for curative treatment. In Figure 4c, this model shows a two-year relative survival benefit of curative treatment (over palliative) for patients classified as good prognosis (61-23 = 38%) that is 12% higher than for patients classified as poor prognosis (41 − 15 = 26%). In contrast, Figure 4d shows 3% less relative benefit to patients classified by staging as good prognosis versus poor prognosis. Due to the variation in patterns of care, as illustrated in Figure 3d, where an example is given of one centre that varies from the group behavior of increasingly aggressive treatment with time, this form of decision support may only be applicable in a subset of network centres, depending on the proportion of patients already treated curatively in a given centre. For patients classified as poor prognosis by the model who received curative treatment the survival rate at two years is 20% less than those predicted as good prognosis by the model, but also treated curatively. These data might help when discussing patient outcomes and decisions and must be balanced with the risk of radiation-induced toxicities which have not been considered here. Any such decision support needs careful clinical evaluation for confidence in use. The current work is simply model validation of a limited-parameter-set model to demonstrate feasibility using NSCLC as the example. Future potential development would be to improve models to optimise model-based stratification and prediction with the inclusion of more parameters and also to apply to other treatment sites and enable comparison of changing treatment techniques. The assessment of relevant patient numbers who meet the eligibility criteria for the two RTOG RCTs that form the basis of the EviQ guidelines highlights the need for access to and the ability to learn from clinical practice datasets, to supplement and expand on trial evidence. Trial-based treatment guidelines provide treatment recommendations for all patients; however for many (65% of patients with Stage III NSCLC RT from this work), there are no directly applicable RCT-based recommendations for their individual treatment. 'Real-world' outcomes are better discussed with the patient during decision-making, rather than only 'optimistic' RCT results

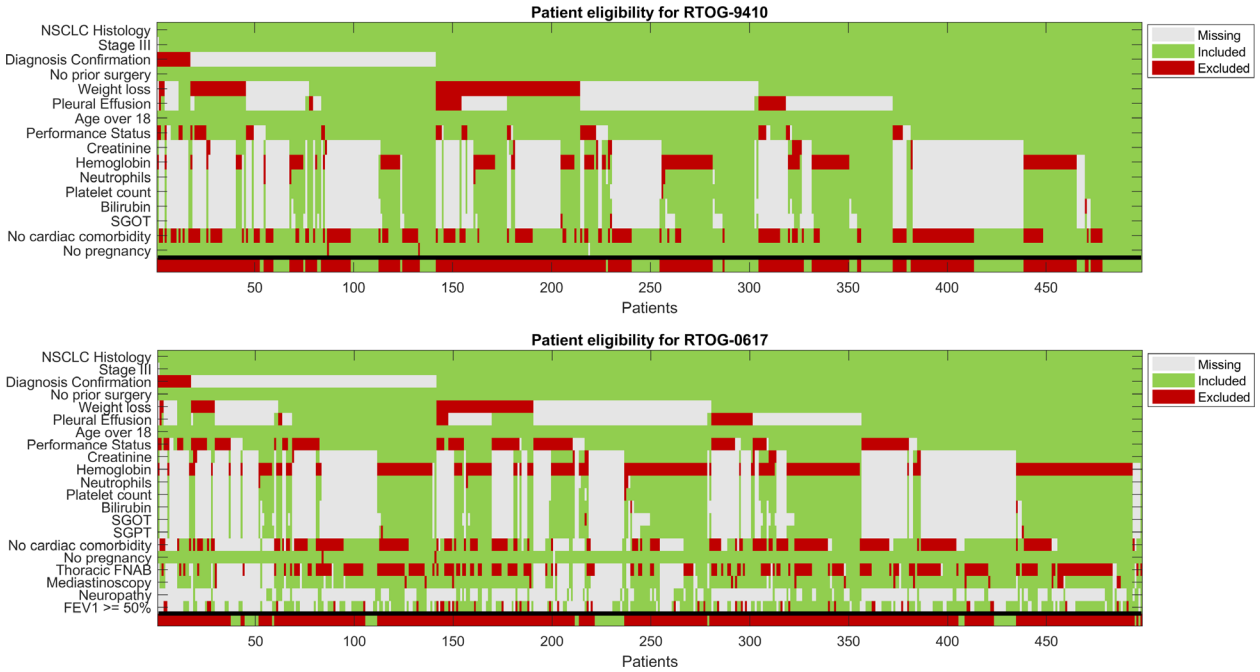

Fig. 5. Data availability graph for Centre A for the eligibility analysis of two RCTs, RTOG-9410 and RTOG-0617. Each row indicates the eligibility for each patient on one criterion. Entries that are green mean the data indicates inclusion whereas red indicates exclusion based on that criteria. Grey entries indicate missing data and the last row indicates the final eligibility per patient under the assumption that missing data, if identified and entered, will indicate inclusion.

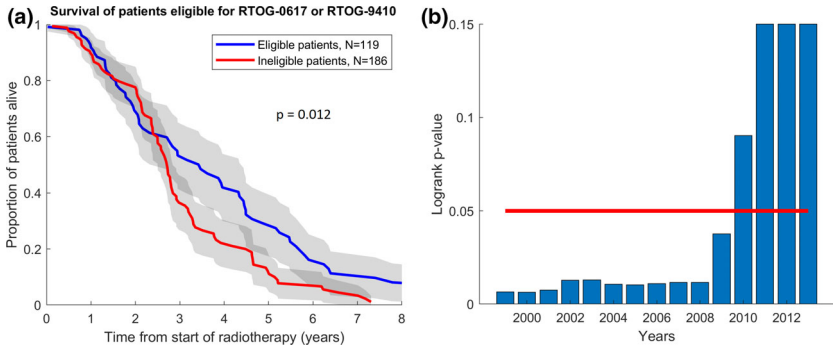

Fig. 6. (a) Survival plot for the curatively treated cohorts of patients meeting the criteria (eligible) and patients not eligible for either of the trials RTOG-9410 or RTOG-0617 treated between the years 2007 and 2019 with 95% confidence intervals shown. (b) Log-rank test applied to cohorts from 2000-2019 up to 2013–2019, where the start year is indicated in the x-axis. Each bar is the *P*-value of the test applied from that year until 2019. The red bar is the threshold of 0.05 significance.

when patients do not meet eligibility criteria and will likely not fully receive the RCT-quoted benefit. Networks such as AusCAT, making clinical practice data available for learning, provide the opportunity to develop additional evidence-based clinical guidelines to support decision making for these patients. In addition, this methodology may enhance the enrolment of patients into future RCTs if decision-support systems highlight those patients that fulfill eligibility criteria.

In future investigations, a similar approach to that used here to assess RCT eligibility could also help to measure the impact of RCT translation into clinical practice and to assess the patient numbers that would be potentially eligible for a proposed RCT. The AusCAT platform could then also support RCT planning and design and provide additional evidence of the impact of RCTs.

The increased use of electronic medical records over time and of automation should improve data completion rates and usability for clinical decision support modelling. For example, novel approaches to collecting data using patient reported outcomes have been shown to help.[27] Anecdotal experience in the network indicates that the

data evaluation and its potential clinical value have focused interest on data quality and completeness and also on resource-efficient ways to improve this. Although promising solutions exist, significant challenges remain in ensuring robust data collection and for streamlined access to data, particularly imaging data. It was clear that structured data for all target variables is not consistently recorded across the centres, although the network's tools can help achieve standard usable formats for its requirements. Potential bias must be clearly assessed when using such data, particularly for patients who have missing data. Models should be developed on variables with standardised definitions that can be practicably incorporated into OIS databases. There are imputation approaches that can be used for managing and substituting some missing data. Previous work within the network showed that mean value imputation had least impact on machine learning models[28]; however it is important to explore these approaches for different datasets and models.

The distributed learning approach to model validation and development using the AusCAT platform enables local datasets to remain secure, overcoming the challenge of moving large datasets (including image datasets) and mitigating some ethics and privacy concerns associated with sharing data, particularly across jurisdictions and countries. Through use of internationally accepted ontologies, the AusCAT network has established links to other international datasets and distributed networks. International collaboration is anticipated to provide opportunities for the Australian community to learn from variation in practice across the globe and from combined larger datasets.

Conclusion

This overview of the AusCAT network demonstrates the feasibility of automatically extracting and standardising clinical practice datasets from multiple Australian centres and a range of clinical systems and making them available for distributed learning for a range of clinical applications and problems, including direct clinical decision support. It highlights the availability of these data for development of additional clinical evidence for patients where RCT evidence may be limited or non-existent. It also allows assessment of research evidence translation into clinical practice and the potential feasibility of proposed future clinical trials. The network is anticipated to expand and to provide opportunities for the radiation oncology clinical and research community to learn from this large body of available data to improve evidence and subsequently patient outcomes.

Acknowledgements

This work was in part supported by a NSW Office of Health and Medical Research (OHMR) Bioinformatics grant, RG14/11, by radiation oncology trust funds from Liverpool and Macarthur Cancer Therapy Centres, Sydney West Radiation Oncology Network, Westmead and Blacktown Hospitals and Illawarra Cancer Care Centre (Wollongong Hospital), by a Hunter Cancer Alliance grant and by Cancer Institute NSW Early Career Fellowship 2019/ECF004.

References

1. Grand MM, O'Brien PC. Obstacles to participation in randomised cancer clinical trials: a systematic review of the literature. *J Med Imaging Radiat Oncol* 2012; **56**: 31–9.
2. Duggan KJ, Descallar J, Vinod SK. Application of guideline recommended treatment in routine clinical practice: a population-based study of stage I-IIIB non-small cell lung cancer. *Clin Oncol* 2016; **28**: 639–47.
3. Myers L, Stevens J. Using EHR to Conduct Outcome and Health Services Research. Secondary Analysis of Electronic Health Records. Springer, Cham (CH), 2016; https://doi.org/10.1007/978-3-319-43742-2_7
4. Wong J, Horwitz MM, Zhou L, Toh S. Using machine learning to identify health outcomes from electronic health record data. *Curr Epidemiol Rep.* 2018; **5**: 331–42.
5. Walpole ET, Theile DE, Philpot S, Youl PH. Development and implementation of a cancer quality index in Queensland, Australia: a tool for monitoring cancer care. *J Oncol Pract* 2019; **15**: e636–43.
6. Price G, van Herk M, Faivre-Finn C. Data mining in oncology: the ukCAT project and the practicalities of working with routine patient data. *Clin Oncol* 2017; **29**: 814–7.
7. Lambin P, Zindler J, Vanneste B *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.
8. Dekker A, Vinod S, Holloway L *et al*. Rapid learning in practice: A lung cancer survival decision support system in routine patient care data. *Radiother Oncol* 2014; **113**: 47–53.
9. Dehing-Oberije C, Yu S, De Ruysscher D *et al*. Development and external validation of prognostic model for 2-year survival of non-small-cell lung cancer patients treated with chemoradiotherapy. *Int J Radiat Oncol Biol Phys* 2009; **74**: 355–62.
10. Lustberg T, Bailey M, Thwaites DI *et al*. Implementation of a rapid learning platform: Predicting 2-year survival in laryngeal carcinoma patients in a clinical setting. *Oncotarget.* 2016; **7**: 37288–96.
11. Jochems A, Deist TM, van Soest J *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.

12. Deist TM, Dankers FJWM, Ojha P *et al*. Distributed learning on 20 000+ lung cancer patients – The Personal Health Train. *Radiother Oncol* 2020; **144**: 189–200.
13. Shi Z, Zhovannik I, Traverso A *et al*. Distributed radiomics as a signature validation study using the Personal Health Train infrastructure. *Sci Data* 2019; **6**: 218.
14. Gaye A, Marcon Y, Isaeva J *et al*. DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int J Epidemiol* 2014; **43**: 1929–44.
15. Lu C-L, Wang S, Ji Z *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.
16. Field M, Barakat MS, Bailey M *et al*. *A Distributed Data Mining Network Infrastructure for Australian Radiotherapy Decision Support*. Engineering and Physical Sciences in Medicine (EPSM), Wellington NZ, 2015; 323.
17. eviQ Cancer Treatments Online. *Respiratory Non-Small Cell Lung Cancer EBRT*. Cancer Institute NSW, 2006. Available from URL: https://www.eviq.org.au/radiation-oncology/respiratory/1871-respiratory-non-small-cell-lung-cancer-ebrt.
18. Alexander M, Wolfe R, Ball D *et al*. Lung cancer prognostic index: a risk score to predict overall survival after the diagnosis of non-small-cell lung cancer. *Br J Cancer* 2017; **117**: 744–51.
19. Jochems A, El-Naqa I, Kessler M *et al*. A prediction model for early death in non-small cell lung cancer patients following curative-intent chemoradiotherapy. *Acta Oncol* 2018; **57**: 226–30.
20. Jayasurya K, Fung G, Yu S *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. https://doi.org/10.1118/1.3352709.
21. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform* 2007; **40**: 30–43.
22. 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**: e854–62.
23. Kalendralis P, Sloep M, van Soest J, Dekker A, Fijten R. Making radiotherapy more efficient with FAIR data. *Phys Med* 2021; **82**: 158–62.
24. Wilkinson MD, Dumontier M, Aalbersberg IJ *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016; **3**: 160018.
25. Curran WJ, Paulus R, Langer CJ *et al*. Sequential vs. concurrent chemoradiation for stage III non-small cell lung cancer: randomized phase III trial RTOG 9410. *J Natl Cancer Inst* 2011; **103**: 1452–60.
26. Bradley JD, Paulus R, Komaki R *et al*. Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial phase 3 study. *Lancet Oncol* 2015; **16**: 187–99.
27. Girgis A, Delaney GP, Arnold A *et al*. Development and feasibility testing of PROMPT-Care, an eHealth system for collection and use of patient-reported outcome measures for personalized treatment and care: a study protocol. *JMIR Res Protoc* 2016; **5**: e227.
28. Barakat MS, Field M, Ghose A *et al*. The effect of imputing missing clinical attribute values on training lung cancer survival prediction model performance. *Health Inf Sci Syst.* 2017; **5**: 16.