BMC Medical Informatics and Decision Making

RESEARCH

Open Access

A series of natural language processing for predicting tumor response evaluation and survival curve from electronic health records

Toshiki Takeuchi[1*], Hidehito Horinouchi[2], Ken Takasawa[3], Masami Mukai[4], Ken Masuda[2], Yuki Shinno[2], Yusuke Okuma[2], Tatsuya Yoshida[2], Yasushi Goto[2], Noboru Yamamoto[2], Yuichiro Ohe[2], Mototaka Miyake[5], Hirokazu Watanabe[5], Masahiko Kusumoto[5], Takashi Aoki[1], Kunihiro Nishimura[1] and Ryuji Hamamoto[3]

Abstract

Background The clinical information housed within unstructured electronic health records (EHRs) has the potential to promote cancer research. The National Cancer Center Hospital (NCCH) is widely recognized as a leading institution for the treatment of thoracic malignancies in Japan. Information on medical treatment, particularly the characteristics of malignant tumors that occur in patients, tumor response evaluation, and adverse events, was compiled into the databases of each NCCH department from EHRs. However, there have been few opportunities for integrated analysis of data on both the hospital and research institute.

Methods We developed a method for predicting tumor response evaluation and survival curves of drug therapy from the EHRs of lung cancer patients using natural language processing. First, we developed a rule-based algorithm to predict treatment duration using a dictionary of anticancer drugs and regimens used for lung cancer treatment. Thereafter, we applied supervised learning to radiology reports during each treatment period and constructed a classification model to predict the tumor response evaluation of anticancer drugs and date when the progressive disease (PD) was determined. The predicted response and PD date can be used to draw a survival curve for the progression-free survival.

Results We used the EHRs of 716 lung cancer treatments at the NCCH and structured data of the cases as labels for the training and testing of supervised learning. The structured data were manually curated by physicians and CRCs. We investigated the results and performance of the proposed method. Individual predictions of tumor response evaluation and PD date were not extremely high. However, the final predicted survival curves were nearly similar to the actual survival curves.

*Correspondence:
Toshiki Takeuchi
take@xcoo.jp

Full list of author information is available at the end of the article

Conclusions Although it is difficult to construct a fully automated system using our method, we believe that it achieves sufficient performance for supporting physicians and CRCs constructing the database and providing clinical information to help researchers find out a chance of clinical studies.

Keywords Electronic health records, Natural language processing, Deep learning, Tumor response evaluation, Survival curve

Introduction

Cancer is one of the major causes of death in the world [1] and particularly the most common cause of death in Japan [2]. The treatment of cancer has advanced significantly as a result of constant basic research, translational research, and clinical research. In clinical research, randomized trials are the most important. However, the results obtained from randomized trials have often diverged from actual clinical practice. In this context, clinical research using real world data (RWD) is attracting attention, since RWD can reflect treatment efficacy and adverse events in actual clinical practice that cannot be obtained in randomized trials. The widespread use of electronic health records (EHRs) is a major driver in making RWD creation possible. However, the data in EHRs themselves are not structured. There is a great deal of interest in how to change unstructured EHR data into a form that can be used for clinical research.

The National Cancer Center Hospital (NCCH) is widely recognized as a leading institution for the treatment of thoracic malignancies in Japan [3]. Information on medical treatment, such as the characteristics of malignant tumors that occur in patients and tumor response evaluation, was individually structured and compiled into the databases of each NCCH department from EHRs. The structured data were used for various studies using the EHRs information on the laboratory side, such as survival analysis. However, there have been few opportunities for integrated analysis of data on both the clinical and laboratory sides. This is because the database systems are managed by each department, and data is not a subset of EHRs and need to be built. In addition, another reason is that EHRs exist within a closed network for the protection of personal information. Physicians and clinical research coordinators (CRC) manually collected and structured the EHRs data, which is a burdensome task. Particularly, the tumor response evaluation of cancer treatment was difficult because they determined the tumor response evaluation comprehensively by reading radiology reports.

In this study, we aimed at developing a supporting tool for conducting studies using the EHRs more efficiently. Specifically, treatment periods, tumor response evaluation, dates that progressive disease (PD) was confirmed, and survival curves are utilized for various clinical studies using the EHRs. An automated system or a suggestion system for providing these information help physicians and CRCs curating the structured database. Note that this study did not intend to affect the clinical decision. We proposed a method comprising a series of analyses for predicting treatment periods, tumor response evaluation, and survival curves of drug therapy for patients with lung cancer from the real EHRs using natural language processing (NLP). The proposed method provided such clinical information to help researchers find out a chance of clinical studies.

Related work

NLP is an instrumental technology for clinical decision support aimed at helping medical workers make decisions [4]. Some medical information, such as impression in the EHRs and interventions of a clinical trial, is unstructured and described in free-text format. Information extraction and structuring from such unformatted data using NLP strongly supports advanced analyses [5, 6].

Unstructured data in clinical studies often contain key-value data. Kumamaru proposed a general approach that can efficiently extract a broad range of numeric human physiological data such as blood pressure, EF, and laboratory values from narrative notes [7]. Cai et al. also developed and published a tool for extracting EHRs numerical data [8]. Their rule-based approach achieved high accuracy for capturing numerical vital sign values. The key-value data are relatively easy to be manipulated by the rule-based algorithm.

Clinical variables in cancer treatment, such as tumor response evaluation, are not directly represented in EHRs. Recently, machine learning (ML) algorithm, particularly deep learning, has received considerable attention because of its performance and flexibility. Liao et al. predicted Crohn's disease using the adaptive LASSO penalized logistic regression [9]. Zhang et al. developed a semi-supervised phenotyping pipeline, PheCAP, using ML approaches that produced the probability of the phenotype [10]. Yuan et al. used an ML algorithm for a lung cancer prognosis [11]. Kehl et al. assessed whether deep natural language processing can extract relevant cancer outcomes from radiologic reports of the English EHRs at the United States [12]. Araki et al. developed models based on Bidirectional Encoder Representations from Transformers (BERT) to extract some outcomes from Japanese EHRs [13].

Although there are some studies that achieve high performance using EHRs and NLP, few studies focus on actual medical analyses in an actual medical field. Overall analysis process using real EHRs is more important for promoting clinical studies using the EHRs. The structured clinical database in actual field included mistakes, spelling inconsistencies, various formats, omissions, and wrong decisions because operations compiling from the EHRs into the database were burdensome tasks.

Methods

Data

In this study, we selected the EHRs of 716 treatments for lung cancer at the Department of Thoracic Oncology, the NCCH. Each treatment had a single regimen consisting of one or more anticancer drugs. Although a patient could have several treatments, the treatments did not performed simultaneously. The 716 treatments information also existed in the structured database of the Department of Thoracic Oncology and was well-reviewed and reliable for evaluating our method. Incomplete cases were excluded in advance, and the EHRs and the structured data of the 716 treatments had no missing data. Patients were aged 24–91 years at enrollment for recruitment from 1998 to 2019. We used injection records, prescription records, and radiology reports in the EHRs. The injection records included a date and an anticancer drug injected into the patient, and the prescription records included a date and an oral anticancer drug. Radiology reports included the date, a method, findings, and impression of a radiation examination.

Additionally, we used structured data of the 716 lung cancer treatments as labels for the training and testing of supervised learning. The structured data were manually curated by physicians and CRCs in the Department of Thoracic Oncology at the NCCH. The structured data comprised patient information, treatment records, and hospital records. We used treatment records that included treatment periods, tumor response evaluation, and dates that progressive disease (PD) was confirmed. The clinical information in this study was determined by physicians and CRCs only for easily reference from clinical studies.

Fig. 1 shows examples of the EHRs and the structured data used in this study. The examples are dummy data similar to the original data because the original data are protected and are not publicly available.

Method overview

Our method comprised three sequential analyses: prediction of treatment periods, prediction of tumor response evaluation and PD date, and survival analysis (Fig. 2). We developed a rule-based algorithm to predict treatment duration using a dictionary of anticancer drugs and regimens used for lung cancer treatment. Thereafter, we applied supervised learning to radiology reports during each treatment period and constructed a classification model to predict the tumor response evaluation of anticancer drugs and the date when the PD was determined. The predicted response and PD date can be used to plot a survival curve for the progression-free survival (PFS) period.

Electronic health records (EHRs)

Injection

PatientNo	DateTime	MedicationName
1	2015-12-08	カルボプラチン（50，150，450mg）

Carboplatin

Prescription

PatientNo	DateTime	MedicationName
1	2018-01-15	タルセバ錠250mg

Tarceva

Radiology report

PatientNo	ExaminDateTime	Method	Findings	Impression
1	2016-02-05	CT	2016年6月1日のCT検査の結果と比較しました。 2016年3月13日に左下葉切除術が施行されています。 前縦隔リンパ節転移は縮小。明らかな肺転移なし。 左少量胸水あり。 肝S5を圧排している脂肪性腫瘤に著変なし。 副腎：有意な腫大なし。リンパ節：有意な腫大なし。 腹水なし。脳転移を疑う所見なし。	縦隔リンパ節転移：縮小 Mediastinal lymph node metastasis: decreased

The results were compared to the June 1, 2016 CT scan results. A left lower lobectomy was performed on March 13, 2016. Anterior mediastinal lymph node metastasis reduced. No obvious lung metastases. Small amount of left pleural effusion. No significant change in fatty mass draining hepatic S5. Adrenal glands: no significant enlargement. Lymph nodes: no significant enlargement. No ascites. No brain metastasis.

Structured data

Treatment record

PatientNo	Treatment detail	Treatment start	Treatment end	Tumor response evaluation	PD date
患者ID	治療詳細	治療開始日	治療終了日	効果	PD確定日
1	Gefitinib	2012-06-15	2014-08-14	PD	2014-03-13

Fig. 1 Examples of the electronic health records (EHRs) and the structured data used in the study. Red texts are translations into English

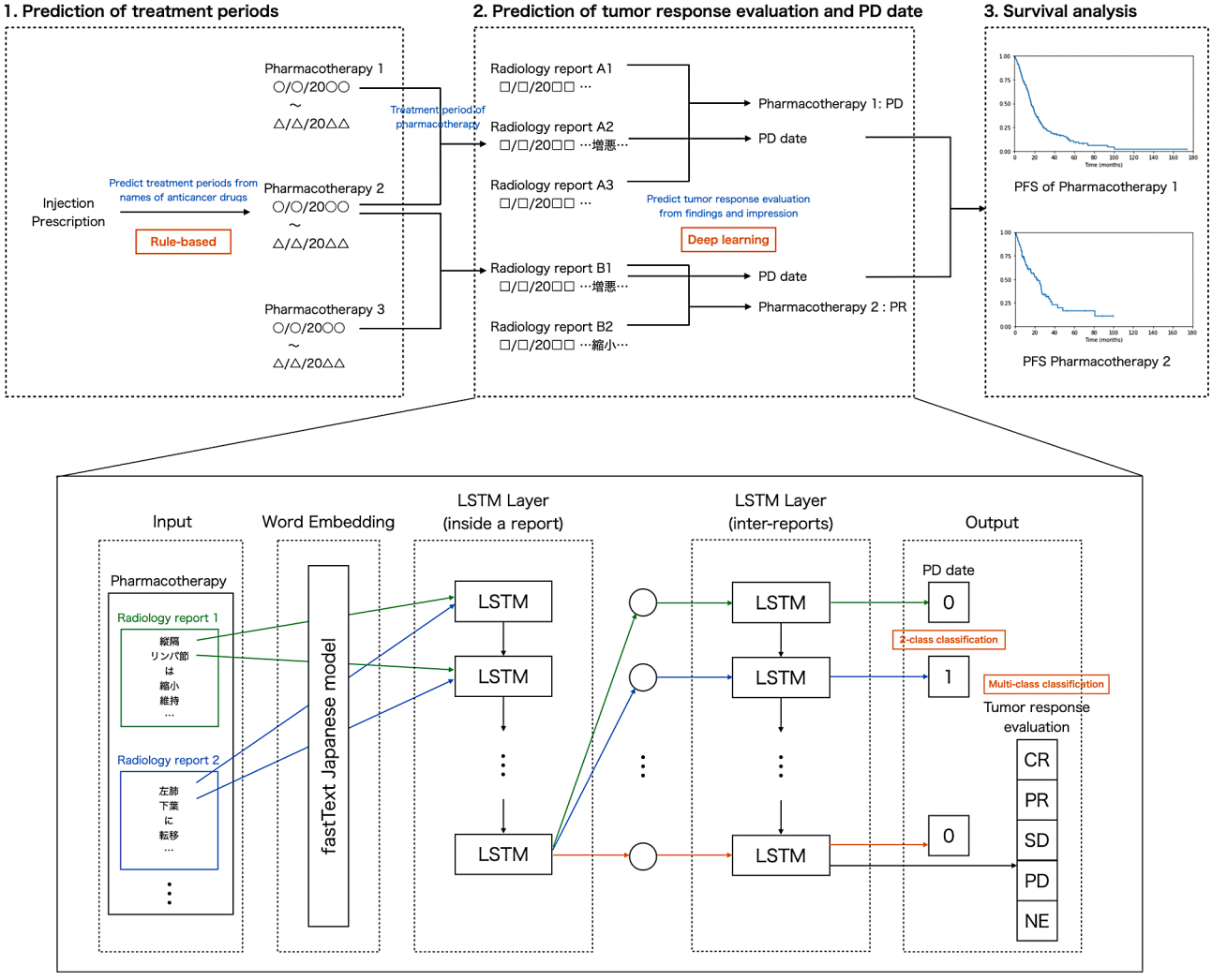

Fig. 2 Overview of the proposed method

Although the EHRs data and the structured data used in this study are protected and are not publicly available, our code used to train and evaluate the models with dummy data is available on GitHub (https://github.com/xcoo/jp-cancer-ehrs-analysis). The EHRs and structured data in the public code are dummy, whereas the ML models were trained with the real EHRs data. Predictions of treatment periods and tumor response evaluation described later are able to be tried with the dummy data.

Rule-based prediction of treatment periods

The duration of treatment with injectable and oral anticancer drugs was predicted from injection and prescription records. The treatment period was defined as the duration between the first injection/prescription and last injection/prescription. In this step, we adopted a rule-based algorithm because people's decisions do not intervene during the treatment period.

Algorithm S1 is pseudocode of the main routine of the rule-based algorithm for predicting the treatment periods. First, we extracted only anticancer drugs from the injection and prescription records (line 3). We prepared and used a dictionary of anticancer drugs based on the NCI Drug Dictionary, DrugBank [14], and Interlanguage links of Wikipedia. Some anticancer drugs are used alone, whereas others are used simultaneously. 42.5% of treatments of our dataset used regimens including multiple anticancer drugs. For instance, gefitinib for lung cancer is mainly used without other anticancer drugs. In contrast, pemetrexed (PEM) is frequently used with other drugs; thus, we grouped the anticancer drugs based on an injection/prescription date (line 12). Additionally, the anticancer drug group was converted into a regimen name, such as CDDP + PEM and CBDCA + PEM, using a regimen dictionary. Finally, we obtained the treatment period consisting of the first and last injection/prescription dates

and the regimen name (line 25). The predicted treatment periods were used for the trailing steps.

Prediction of tumor response evaluation and PD date using machine learning

The tumor response evaluation comprises complete response (CR), partial response (PR), stable disease (SD), and PD [15]. The tumor response evaluation in this study was assigned to each clinical treatment, and individual evaluation of each radiology report was not provided. Among the four responses, PD is the most important response in clinical treatment because it provides important information for determining the next treatment, and the date of the PD is frequently used for survival analysis. The data in this study used not evaluable (NE) in addition to the four responses. The percentage of each response in our dataset was CR 1.2%, PR 30.0%, SD 28.8%, PD 20.0%, and NE 20.0%. CR cases are extremely rare in lung cancers and are insufficient for training a model. We used the four responses except for CR (that is, PR, SD, PD, and NE) and the date that the PD of a treatment was confirmed.

Fig. 2 contains the full neural network constructed in this study. The input is a series of radiology reports related to a pharmacotherapy treatment. A radiology report includes an examination date, a locus, a diagnosis, a purpose, findings, an impression, and other miscellaneous information. We particularly selected the findings and the impression as input to our model because clues to evaluate the tumor response evaluation such as changes of tumor size and tumor spreads were found in the findings and the impression. Each input free text is conjunction of findings and impression of the radiology report. The input text is tokenized using a Japanese tokenizer included in Stanza [16]. The tokens were converted into word-embedding representations of fastText [17]. We rebuilt and used the Japanese model of fastText with Stanza tokenizer.

The characteristics of the network is two different Long Short-Term Memory (LSTM) layers [18, 19]. LSTM is a recurrent neural network (RNN) architecture that has been successfully used for sequence prediction tasks. LSTM is frequently used for NLP because texts written in a natural language can be treated as a sequence of words or characters. We used PyTorch [20] to implement the neural network. The first LSTM layer considers a single free text in the radiology report and the second considers a sequence of multiple radiology reports. The number of features in the hidden state was set to 32 for the first LSTM and 16 for the second LSTM. The log-softmax function is used as the activation function.

The network finally outputs two different predictions: tumor response evaluation and PD date. We treated the prediction of tumor response evaluation as a multiclass classification problem and that of PD date as multiple binary classifications. This problem design is similar to many-to-one model of LSTM resolving the text classification [21] and many-to-many model of that resolving the Part-of-Speech (POS) tagging [22]. We applied a linear layer to the final hidden states of the second LSTM layer for transforming the states to four classes, that is, 0 (PR), 1 (SD), 2 (PD), and 3 (NE). We applied another linear layer to a hidden state of each step and transformed the state to two classes, that is, 0 (not PD) and 1 (PD).

Additionally, we adopted multi-task learning (MTL) [23] to improve performance. MTL is known as a method to leverage the related and useful information contained in multiple learning tasks. In our method, tumor response evaluation and PD date were simultaneously trained.

Survival analysis using predicted data

PFS was defined as the time from randomization or initiation of treatment to the occurrence of disease progression or death. PFS is widely used as a surrogate endpoint in oncological clinical trials [24–26]. We plotted Kaplan-Meier curves [27] of PFS from the PD date predicted by the method in the Prediction of Tumor Response Evaluation and PD Date Using Machine Learning.

An event of PFS is the date on which progression or death is detected, and censoring of PFS is the end date of the treatment or the last follow-up date [28]. We used our PD date prediction as the progression event and referred to the structured database to obtain the death date and the last follow-up date. We used a Python library, lifelines [29], to plot the Kaplan-Meier curves.

Results

We split the actual EHRs of 716 lung cancer treatments into 573 treatments (80%) for training and 143 treatments (20%) for testing. Moreover, we used 10% of training data for parameter optimization using grid search. Although some ML studies additionally prepared validation data for early stopping [30], we used fixed epochs because the EHRs data were not sufficiently large in this study. We located a dropout layer just after the second LSTM layer to prevent overfitting during training. We used the negative log likelihood loss for loss function and RAdam for optimization. ML model training and statistical analyses were performed using Python v3.9.13, CUDA v11.8, PyTorch v2.4.0, stanza v1.4.2, scikit-learn v1.5.0, and lifelines v0.27.3.

Treatment periods

Fig. 3(a) shows the accuracy of the predicted treatment periods. We used treatment periods in the structured data as a baseline. The accuracy of the prediction corresponding to the ground truth is 0.564. The accuracy

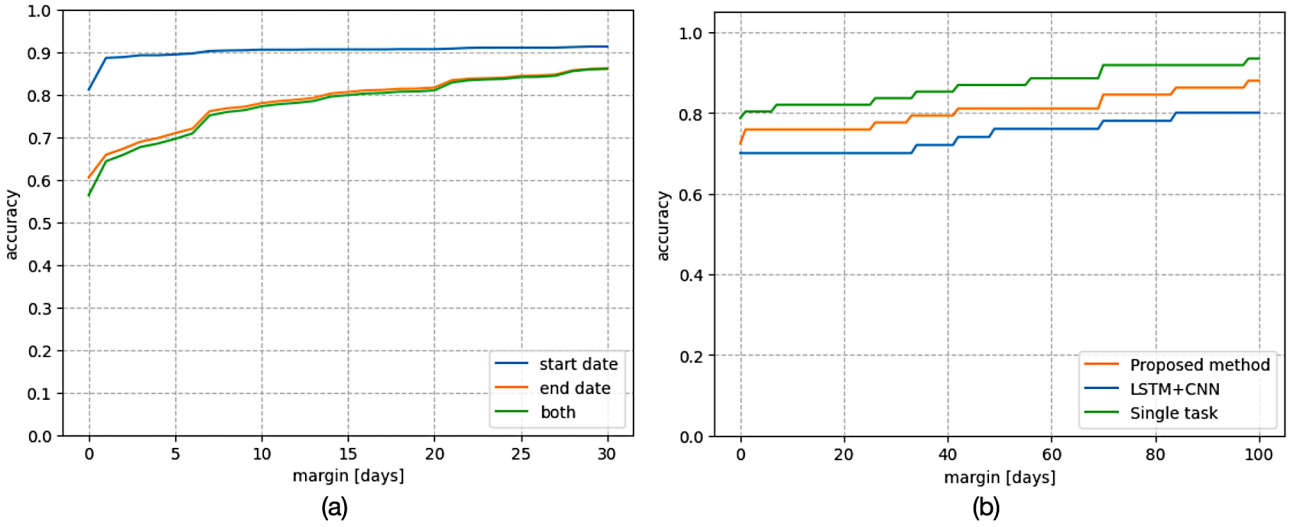

Fig. 3 Accuracy of (**a**) the predicted treatment periods and (**b**) the predicted PD dates. The x-axis shows approval margins within which the predicted date is considered as correct if the difference from an answer date is within the margin days

Table 1 Precision, recall, and f1-score of the prediction of tumor response evaluation

	precision	recall	f1-score
PR	0.77	0.82	0.80
PD	0.68	0.71	0.69
SD	0.67	0.62	0.64
NE	0.75	0.75	0.75

increased to 0.644 with a one-day margin, 0.796 with a two-weeks margin, and 0.862 with a one-month margin. The accuracy of the start date was 0.887, with only a one-day margin. The reason for significant increase with a one-day margin is that we cannot detect a patient taking medicine, whether on a prescription date or not, from the prescription records. The start date of the treatment period is easily predicted because it is only an injection date or the day following a prescription in most cases. However, the precise end date of the treatment period is uncertain because the patient's last dose is not logged by a hospital. To obtain a precise end date, we must refer to information from a medical interview. In this study, however, the following steps use only the start date of each treatment period; thus, the accuracy of 0.9 of the start date prediction is sufficient.

Tumor response evaluation

Table 1 lists the precision, recall, and f1-score for the prediction of tumor response evaluation. The performance of PR was highest and that of SD was lowest. The former reason was that texts of PR case probably tended to include specific phrases such as "decrease." However, such specific phrases were not included in the texts of SD case. The micro averages of the precision, recall, and f1-score were 0.72, which was equivalent to accuracy. The macro averages were also 0.72 in this case.

Next, we compared our method with two other methods: LSTM + CNN and a single task. LSTM + CNN is a method in which the second LSTM layer in Fig. 2 is changed to a convolutional neural network (CNN) layer [31]. CNN is not an RNN architecture and is not as strong for sequence prediction tasks as LSTM. We used the CNN architecture proposed by Zhang et al. [32] instead of the second LSTM layer. In contrast, a single task is a method that trains tumor response evaluation and PD date individually. We attempted to confirm the sequence relation in radiology reports by comparing our method to LSTM + CNN and the task relation by comparing our method to a single task.

Table 2 summarizes the results of the comparison between LSTM + CNN and a single task. Comparing the micro/macro averages, the proposed method showed higher performance than the two other methods. Particularly, the performance of LSTM + CNN was lowest; thus, we found that the second LSTM layer in Fig. 2 made a significant contribution to the prediction of tumor response evaluation. From the result of single task, MTL also increased the performance in the prediction of tumor response evaluation.

Table 2 Precision/Recall/F1-scores of the predictions of tumor response evaluation

	Proposed method	LSTM+CNN	Single task
PR	0.77/**0.82**/**0.80**	**0.80**/0.80/**0.80**	0.73/0.73/0.73
PD	**0.68**/**0.71**/**0.69**	0.45/0.58/0.51	0.56/0.62/0.59
SD	0.67/0.62/0.64	0.61/0.53/0.57	**0.70**/**0.66**/**0.68**
NE	**0.75**/**0.75**/**0.75**	0.68/0.65/0.67	0.60/0.60/0.60
micro avg	**0.72**	0.65	0.67
macro avg	**0.72**/**0.72**/**0.72**	0.64/0.64/0.64	0.65/0.65/0.65

PD date

Fig. 3(b) shows a plot of the accuracy of the predicted PD dates. The accuracy of the proposed method in which the prediction corresponds only to the ground truth is 0.728. The accuracy increased to 0.771 with a one-day margin, 0.818 with 43 days margin, and 0.882 with 100 days margin. The long margins such as 43 days and 100 days are permissible for speedup of survival analysis described later because the time scale of survival analysis is months or years. In comparison with LSTM + CNN and a single task, the accuracy of LSTM + CNN is lower and that of the model learned with a single task is higher than that of our method. The effect of MTL on the PD date prediction conflicted with that of the tumor response evaluation prediction. We chose different weights of loss functions of the tumor response evaluation task and PD date task for balancing performance of each task in MTL. Consequently, the weights were a bit more optimized for the tumor response evaluation task than the PD date task. We thought the balance of weights was better because a bit of degradation of the PD date prediction did not severely affect the following survival analysis.

Progression-free survival curves

Fig. 4 shows survival curves of PFS for all treatments and each regimen. We selected the top six regimens in the order of the number of patients because regimens with few treatment records were not reliable. The PFS curves of CDDP + VNR, CDDP + PEM and CBDCA + PEM were significantly precise; however, those of gefitinib and erlotinib were different when using the actual PD date and the predicted PD date. Although the PFS curves of PEM relatively corresponded within a year, there was a significant difference after more than a year. The PFS curves of cytotoxic anticancer drugs tend to be highly accurate, and those of molecularly targeted drugs tend to have low-accuracy. According to a survey of our dataset, 67.9% of treatments used only cytotoxic anticancer drugs and 32.1% of treatments included molecularly targeted drugs. The molecularly targeted drugs have become popular since around 2010 in Japan and thus, data of the molecularly targeted drugs were small compared to the cytotoxic cancer drugs. We think PFS prediction of the molecularly targeted drugs can be improved as more treatments using the molecularly targeted drugs are conducted in the future.

We also performed a survival analysis for all treatments. Although the two PFS curves are nearly similar, they are slightly different from 15 months to 35 months and from 40 months to 60 months. The median survival time of the predicted PFS was approximately two months longer than that of the actual PFS.

Discussion

Individual predictions of tumor response evaluation and PD date were not extremely high. We think the performance limitation is mainly caused by small data size. Although 573 training data are relatively small to train a deep neural network model, the EHR data are not easily available due to problems of private information and patient consent. A joint research of multiple medical institutions would be ideal to collect large dataset of EHRs. However, the final predicted PFS curves were nearly similar to the actual PFS curves. Although it is difficult to construct a fully automated system using our method, our results show performance for supporting the structured database creation aimed at efficient clinical studies. The tumor response evaluation and the detection of PD date are difficult and burdensome tasks even for physicians and CRCs because they need to refer to both complicated texts in a radiology report and RECIST guideline. Displaying predictions of our method quickly provides them first impression of a treatment case. Although the PFS analysis requires complete treatment periods and PD dates, our method can generate PFS curves roughly without waiting for curation of the structured database.

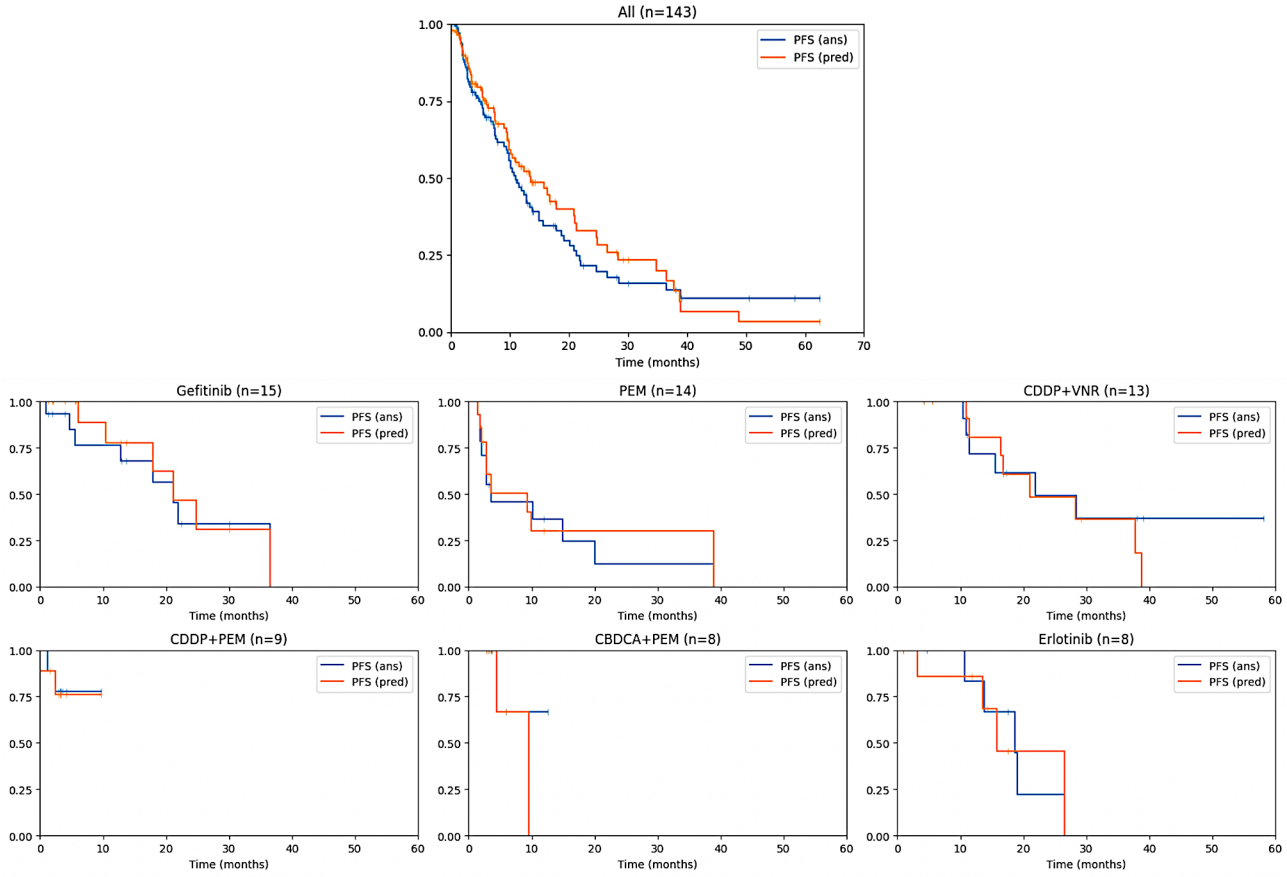

Fig. 4 PFS curves of all treatments and treatments by regimen. The orange line denotes the PFS curve drawn with the predicted PD dates, and the blue line denotes the PFS curve drawn with the actual PD dates. The term *n* next to the title denotes the number of patients to whom the treatment is administered

Figures S2 and S3 show the findings and impression of the radiology report highlighted by integrated gradients (IG) algorithm [33]. IG is a technique that explains the relationship between a model's predictions and its features. The tumor response evaluation in the ground truth and predicted tumor response evaluation in Figure S2 are both PR. The words related to PR in the radiology report A1 and A2, such as 縮小 (*decrease in size*) and 腫大なし (*no evidence of cancerous lesion*), are strongly highlighted, indicating that our model was correctly trained. In contrast, though the tumor response evaluation in the ground truth of Figure S3 is the PD, the predicted tumor response evaluation of Figure S3 is the PR. Words related to PD, such as 増大 (*increase in size*), are strongly highlighted in the radiology report B3. However, the words related to PR, such as 縮小 (*decrease in size*) and 縮小を維持 (*maintained decreased size*), are found in the radiology report B1 and B2. Our model detected tumor response evaluation as PR based on the latter words. Figure S3 is a complex case, and it is difficult to evaluate the response even for physicians and CRCs who created the structured database.

Although each analysis step in our method can be improved or replaced with other models, the sequential analyses are essential to create structured database for clinical studies. The real-world EHRs include dirty data such as incomplete sentences and meaningless copies, and there are no public EHR datasets reflecting such dirty data. Therefore, using real-world EHRs is important to develop a method for real clinical environment. We think future studies of this study should focus on evaluating utilization of our method in real clinical research environment instead of performance improvement on the method. Our method compared to recent general models, such as large language models (LLMs), is naive but easy to introduce in terms of computer hardware cost. Spread of the lightweight AI in clinical environments is also important to promote clinical studies. The training/testing datasets used in this study had the same information categories, and neither data leakage nor target leakage occurred. Input features used in our model such as findings and impression in a radiology report are common for other types of cancer or at the same clinical department of other hospitals. Thus there is no data

leakage between our model and such external data, and we can adopt our model as it is to the EHRs of patients of other cancers or the same department of other hospitals if the data are available. However, there were no appropriate external validation datasets that were easily available, had identical clinical information, and were written in the same language. Validation of our model and confirmation of the model robustness by using the EHRs of other hospitals in the future research indicates spread of our model in real clinical research environment. However, the diseases except for cancer and other departments of hospital might not be able to use our model because the input data are different.

The rapid evolution of AI-enhanced medical decision support systems require ethical considerations, including patient privacy, data security, transparency, accountability, fairness, and bias mitigation [34]. Moreover, in recent years, LLMs have been widely spread over various domains counting medicine, and concerns that the LLMs could disrupt trust factors like factual consistency and process transparency were raised [35]. Utilization environment of our method is supposed to be a research department of a hospital; thus our method does not cause patient-side ethical considerations like patient privacy and data security. Our method used the rule-based algorithm and the LSTM-based model, and the model outputs were fixed values. Consequently, our method does not generate violation of the factual consistency like hallucination. In contrast, the LSTM-based model is disadvantageous for the transparency and the accountability. Although the end-user of our method is a medical researcher, the acceptability of the AI prediction is also an issue. The model interpretability techniques like IG algorithm used for visualizing the relationship between the model's predictions and words in a radiology report is useful for understanding the model behavior.

Conclusions

We developed a method comprising three sequential analyses: a rule-based algorithm to predict treatment duration using a dictionary of anticancer drugs and regimens used for lung cancer treatment, a classification model to predict the tumor response evaluation and the date when the PD was determined, and survival analysis of PFS. We implemented the method with the data of the corresponding cases from the Thoracic Oncology database of the NCCH. We believe that the accuracy of each analysis step is not extremely high, and that the final PFS curve can be used for supporting researchers conduct studies using EHRs.

The usefulness of our method in actual practices and robustness to data of other clinical affiliations have not been confirmed. Input features used in our model such as findings and impression in a radiology report are common for other types of cancer or at the same clinical department of other hospitals. We hope to adapt our method to the EHRs of other cancers and other hospitals to confirm its robustness.

Abbreviations

CNN	Convolutional Neural Network
CR	Complete Response
CRC	Clinical Research Coordinators
EHRs	Electronic Health Records
IG	Integrated Gradients
LSTM	Long Short-Term Memory
ML	Machine Learning
MTL	Multi-Task Learning
NCCH	National Cancer Center Hospital
NE	Not Evaluable
NLP	Natural Language Processing
PEM	Pemetrexed
POS	Part-of-Speech
PR	Partial Response
PD	Progressive Disease
PFS	Progression-Free Survival
RWD	Real World Data
RNN	Recurrent Neural Network
SD	Stable Disease

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-02928-6.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Acknowledgements

Not applicable.

Author contributions

TT developed the methodology, wrote the software, conducted the investigation, and wrote the manuscript. HH conceptualized the study and reviewed/edited the manuscript. KT developed/validated the methodology and reviewed/edited the manuscript. MM curated the resources and the data. KM, YS, YO, and TY curated the data. YG and NY curated the data and reviewed/edited the manuscript. YO curated the data curation. MM and HW curated the data and reviewed/edited the manuscript. MK curated the data. TA reviewed/edited the manuscript. KN reviewed/edited the manuscript and supervised the study. RH conceptualized the study, reviewed/edited the manuscript, and administered the project. All authors read and approved the final manuscript.

Funding

This work was supported by AMED Innovative Cancer Medical Practice Research Project (Grant Number JP22ck0106643), JST CREST (Grant Number JPMJCR1689), JSPS Grant-in-Aid for Scientific Research on Innovative Areas (Grant Number JP18H04908), JST AIP-PRISM (Grant Number JPMJCR18Y4), and MEXT subsidy for Advanced Integrated Intelligence Platform to R.H.

Data availability

Data cannot be shared publicly due to patient confidentiality. The data underlying the results presented in the study are available from the National Cancer Center Japan for researchers who meet the criteria for access to confidential data. Our code used to train and evaluate the models with dummy data is available on GitHub (https://github.com/xcoo/jp-cancer-ehrs-analysis).

Declarations

Ethics approval and consent to participate

This study was approved by the local medical ethical committee at the National Cancer Center Japan (number 2019 – 251). As the opt-out consent process for use of the data from electronic health records was granted under the ethical guidelines for the Medical and Health Research Involving Human Subjects by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), and the Ministry of Health, Labor and Welfare (MHLW) (https://www.mhlw.go.jp/content/10600000/000757206.pdf), explicit patient consent for use of the data was deemed to not be required, and was covered within the National Cancer Center Japan approval (number 2019 – 251).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Xcoo, Inc. Hongo-Sanchome TH Bldg., 6F, 2-40-8, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[2]Department of Thoracic Oncology, National Cancer Center Hospital, 5-1-1 Tsukiji, Chuo-Ku, Tokyo 104-0045, Japan
[3]Division of Medical AI Research and Development, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan
[4]Division of Medical Informatics, National Cancer Center Hospital, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan
[5]Department of Diagnostic Radiology, National Cancer Center Hospital, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan

Received: 11 July 2024 / Accepted: 11 February 2025
Published online: 17 February 2025

References

1. The top 10 causes of death. World Health Organization. 2020. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed 2 Jul 2024.
2. Vital statistics of Japan. 2021. Ministry of Health, Labour and Welfare, 2021. https://www.mhlw.go.jp/toukei/saikin/hw/jinkou/geppo/nengai21/dl/gaikyouR3.pdf. Accessed 2 Jul 2024.
3. Department of thoracic oncology, annual report 2020. The National Cancer Center Hospital. 2020. https://www.ncc.go.jp/en/publication_report/2020/ncch/ncch09.html. Accessed 2 Jul 2024.
4. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009;42(5):760–72.
5. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: from rule-based definitions to machine learning models. Annual Rev Biomedical data Sci. 2018;1(1):53–68.
6. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural language processing for EHR-based computational phenotyping. IEEE/ACM Trans Comput Biol Bioinf. 2018;16(1):139–53.
7. Kumamaru KK, Saboo SS, Aghayev A, Cai P, Quesada CG, George E, Hussain Z, Cai T, Rybicki FJ. CT pulmonary angiography-based scoring system to predict the prognosis of acute pulmonary embolism. J Cardiovasc Comput Tomogr. 2016;10(6):473–9.
8. Cai T, Zhang L, Yang N, Kumamaru KK, Rybicki FJ, Cai T, Liao KP. EXTraction of EMR numerical data: an efficient and generalizable tool to EXTEND clinical research. BMC Med Inf Decis Mak. 2019;19:1–7.
9. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, Gainer VS, Shaw SY, Xia Z, Szolovits P, Churchill S. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ. 2015;350.
10. Zhang Y, Cai T, Yu S, Cho K, Hong C, Sun J, Huang J, Ho YL, Ananthakrishnan AN, Xia Z, Shaw SY. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nat Protoc. 2019;14(12):3426–44.
11. Yuan Q, Cai T, Hong C, Du M, Johnson BE, Lanuti M, Cai T, Christiani DC. Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer. JAMA Netw Open. 2021;4(7):e2114723.
12. Kehl KL, Elmarakeby H, Nishino M, Van Allen EM, Lepisto EM, Hassett MJ, Johnson BE, Schrag D. Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. JAMA Oncol. 2019;5(10):1421–9.
13. Araki K, Matsumoto N, Togo K, Yonemoto N, Ohki E, Xu L, Hasegawa Y, Satoh D, Takemoto R, Miyazaki T. Developing artificial intelligence models for extracting oncologic outcomes from Japanese electronic health records. Adv Therapy. 2023;40(3):934–50.
14. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34(suppl1):D668–72.
15. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45(2):228–47.
16. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. Stanza: a Python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082. 2020.
17. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Association Comput Linguistics. 2017;5:135–46.
18. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
19. Sak H, Senior A, Beaufays F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. arXiv preprint arXiv:1402.1128. 2014.
20. Paszke A, Gross S, Massa F, Lerer A, Bradbury JP, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L. An imperative style, high-performance deep learning library. Adv Neural Inf Process Syst. 1912;32:8026.
21. Rao A, Spasojevic N. Actionable and political text classification using word embeddings and LSTM. arXiv preprint arXiv:1607.02501. 2016.
22. Plank B, Søgaard A, Goldberg Y. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. arXiv preprint arXiv:1604.05529. 2016.
23. Caruana R. Multitask learning. Mach Learn. 1997;28:41–75.
24. Saad ED, Katz A, Hoff PM, Buyse M. Progression-free survival as surrogate and as true end point: insights from the breast and colorectal cancer literature. Ann Oncol. 2010;21(1):7–12.
25. Gill S, Berry S, Biagi J, Butts C, Buyse M, Chen E, Jonker D, Mărginean C, Samson B, Stewart J, Thirlwell M. Progression-free survival as a primary endpoint in clinical trials of metastatic colorectal cancer. Curr Oncol. 2011;18(s2):5–10.
26. Hotte SJ, Bjarnason GA, Heng DY, Jewett MA, Kapoor A, Kollmannsberger C, Maroun J, Mayhew LA, North S, Reaume MN, Ruether JD. Progression-free survival as a clinical trial endpoint in advanced renal cell carcinoma. Curr Oncol. 2011;18(s2):11–9.
27. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457–81.
28. Clinical trial endpoints for the approval of non-small cell lung cancer drugs and biologics. U.S. Department of Health and Human Services Food and Drug Administration. 2015. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-endpoints-approval-non-small-cell-lung-cancer-drugs-and-biologics. Accessed 2 Jul 2024.
29. Davidson-Pilon C. Lifelines: survival analysis in Python. J Open Source Softw. 2019;4(40):1317.
30. Prechelt L. Early stopping-but when?. In Neural networks: tricks of the trade. 2002 (pp. 55–69). Berlin, Heidelberg: Springer Berlin Heidelberg.
31. O'shea K, Nash R. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458. 2015.
32. Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820. 2015.
33. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In International conference on machine learning 2017 Jul 17 (pp. 3319–3328). PMLR.
34. Olorunsogo T, Adenyi AO, Okolo CA, Babawarun O. Ethical considerations in AI-enhanced medical decision support systems: a review. World J Adv Eng Technol Sci. 2024;11(1):329–36.

35. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). NPJ digital medicine. 2024;7(1):183.

Publisher's note