Original Article

Efficient and Reliable Data Extraction in Radiation Oncology using Python Programming Language: A Pilot Study

Rohit Singh Chauhan[1,2], Anirudh Pradhan[3], Anusheel Munshi[2], Bidhu Kalyan Mohanti[4]

[1]Department of Physics, GLA University, [3]Centre for Cosmology, Astrophysics and Space Science, GLA University, Mathura, Uttar Pradesh, [2]Department of Radiation Oncology, Manipal Hospitals, Dwarka, New Delhi, [4]KIMS Cancer Centre, Kalinga Institute of Medical Sciences, KIIT University, Bhubaneswar, Odisha, India

Abstract

Background and Purpose: In recent years, data science approaches have entered health-care systems such as radiology, pathology, and radiation oncology. In our pilot study, we developed an automated data mining approach to extract data from a treatment planning system (TPS) with high speed, maximum accuracy, and little human interaction. We compared the amount of time required for manual data extraction versus the automated data mining technique. **Materials and Methods:** A Python programming script was created to extract specified parameters and features pertaining to patients and treatment (a total of 25 features) from TPS. We successfully implemented automation in data mining, utilizing the application programming interface environment provided by the external beam radiation therapy equipment provider for the whole group of patients who were accepted for treatment. **Results:** This in-house Python-based script extracted selected features for 427 patients in 0.28 ± 0.03 min with 100% accuracy at an astonishing rate of 0.04 s/plan. Comparatively, manual extraction of 25 parameters took an average of 4.5 ± 0.33 min/plan, along with associated transcriptional and transpositional errors and missing data information. This new approach turned out to be 6850 times faster than the conventional approach. Manual feature extraction time increased by a factor of nearly 2.5 if we doubled the number of features extracted, whereas for the Python script, it increased by a factor of just 1.15. **Conclusion:** We conclude that our in-house developed Python script can extract plan data from TPS at a far higher speed (>6000 times) and with the best possible accuracy compared to manual data extraction.

Keywords: Data mining, programming languages, radiotherapy, software, time management

Received on: 31-01-2023 **Review completed on: 27-02-2023** **Accepted on: 02-03-2023** **Published on: 18-04-2023**

Introduction

Data mining in radiotherapy (RT) has always been a challenge for the radiation oncology (RO) community. Difficulty in data extraction arises predominantly because of the poor capturing and handling of the data. To overcome the existing hurdles, it is proposed for a cultural change (manual to automation) in clinical practice, as well as customization of technology, for capturing high-quality treatment data in routine care in a prospective manner.[1] RO makes extensive use of recorded data with transactional value. In recent years, there has been growing interest in capturing relevant RO data to leverage the improvements in several key areas, for example, (a) registries, (b) quality measurement, (c) patient safety, (d) patient care interoperability, and (e) clinical trials and research.[2]

Data mining in RO comprises efforts toward the utilization of informatics for transactional databases, including the approaches to integrate RO data with multimodality databases for a better understanding of the diseases and treatment outcomes.[3,4] To include several hospitals and institutions, across the UK and in Europe, for sharing and mining routine patient data in oncology, a novel distribution learning approach has been attempted by researchers and institutions, keeping in mind the confidentiality and legality involved with the patient data.[5] Machine learning in oncology recognizes the value of routine patient data for "rapid learning" and qualitative measurements as an alternative to laborious randomized control trials. Researchers have recently demonstrated the potential of data mining in selecting Intensity Modulated

Address for correspondence: Mr. Rohit Singh Chauhan, Department of Physics, GLA University, Mathura, Uttar Pradesh, India.
E-mail: rohit.chauhan0011@gmail.com

Access this article online

Quick Response Code:

Website: www.jmp.org.in

DOI: 10.4103/jmp.jmp_12_23

How to cite this article: Chauhan RS, Pradhan A, Munshi A, Mohanti BK. Efficient and reliable data extraction in radiation oncology using python programming language: A pilot study. J Med Phys 2023;48:13-8.

Radiation Therapy (IMRT) beam angles, external beam radiation therapy techniques, and predicting cancer risks.[6-8] Integrated research with data mining and predictive toxicity, modeling can be expanded to personalized patient treatment decision support in RO.[9]

Conventionally, data extraction is done manually, which is laborious, requires human intervention, and is prone to errors. It has been highlighted that, given the extent and complexity of information and records gathered for clinical service, the task before the RO community is made further complicated in having to deal with different vendors, different standards, and different institutional policies as regards the formatting and recording of data.[10] To overcome these challenges, the authors demonstrated a data warehouse project for faster and more efficient data aggregation from different sources within RO.[11]

Since the 1980s, vendor-specific or freestanding treatment planning system (TPS) has been ubiquitous in RO practices. Eclipse (Varian Medical Systems, Palo Alto, CA, USA) has an integrated application programming interface (API). The Eclipse TPS is part of Varian's ARIA, a RO information system (ROIS), and both operate on a single database that stores all patient-related information and treatment records. The technological advances, multiplicity of techniques, and data-intensive treatment processes have combinedly and brought into focus the role of big data and machine learning from RO workflow prospective.[12] To this end and to initiate a transitional environment, we created a novel and faster data mining approach using the vendor-specific API, i.e., PyESAPI (Varian Medical Systems, Palo Alto, CA, USA), environment to fetch desired data from the patients' database of Varian Eclipse TPS with minimum human interventions. Python Eclipse Scripting API (PyESAPI) is a research project that integrates the power of Python with the Varian API ecosystem.[13] With Python, we can use powerful libraries such as NumPy, SciPy, Matplotlib, scikit-learn, pandas, and TensorFlow.

In this pilot study, we want to see how effective and efficient the automated data mining method PyESAPI is compared to the inter-observer variable manual data extraction method that has been used in RO for a long time. "Manual data extraction" refers to the process of collecting or retrieving data from a source or document using human effort and without the use of automated tools or software.

Our hypothesis is that the PyESAPI automated data mining approach will be more efficient and accurate in extracting relevant data from the patients' database of Varian Eclipse TPS, resulting in improved data quality, and reduced human errors compared to the manual data extraction technique. This study aims to demonstrate the potential of PyESAPI in automating the data extraction process in RO and facilitating the utilization of big data for clinical practice and research.

Materials and Methods

Materials

The use of RT in cancer treatment has increased significantly in recent years. RT TPSs play a critical role in the delivery of accurate and safe RT. The Varian TrueBeam linear accelerator and Varian Eclipse TPS are commonly used in RO facilities. These systems share a common patient database called ARIA™. The Eclipse TPS comes with preinstalled software called an API that allows for the creation of scripts to run specific tasks such as reporting plan parameters.

The primary objective of this pilot study is to develop a feasible automation and data mining strategy utilizing existing RO systems, the PyESAPI environment, and the ARIA database of patient records. The secondary objective is to compare the in-house developed data mining strategy with the current manual data extraction practice.

This pilot study was done at a single RO facility with the Varian TrueBeam linear accelerator version 2.7 and the Varian Eclipse Training Box version 15.6. The TBOX system specifications include a 64-bit operating system of either Windows Server 2012 R2 (standard or data center) or Windows Server 2016 (standard or data center). It requires a minimum of 10 vCPU cores and 48 GB of RAM, with an Intel Xeon Silver 4114 at 2.1 GHz or equivalent.

The PyESAPI environment was developed, and the Radiation Oncology data mining script (RODAMS) was written in Python. Figure 1 depicts the alternative path for data extraction utilizing an automated method. The research involved two phases. During Phase I, we compared the speed and accuracy of data extraction using the RODAMS versus the manual extraction technique for a fixed set of 25 features, as shown in Table 1. In Phase II, we examined how increasing the number

Table 1: List of parameters selected for extraction using manual and Radiation Oncology Data Mining Script techniques in phase-I

Hospital information	Patient information	Course details	Diagnosis details	Prescription dose details	Plan details
Hospital name	Patient ID	Course ID	Diagnosis code	Prescribed dose	Plan ID
Hospital location	First name	Course intent	Clinical description	Dose per fraction	Plan intent
	Last name	Starting date		Number of fraction	Treatment orientation
	Sex	Completion date		Target volume ID	Plan normalization method
	DOB	Primary oncologist ID		Treatment approval date	Plan approver name
					Use gating?

DOB: Data of birth

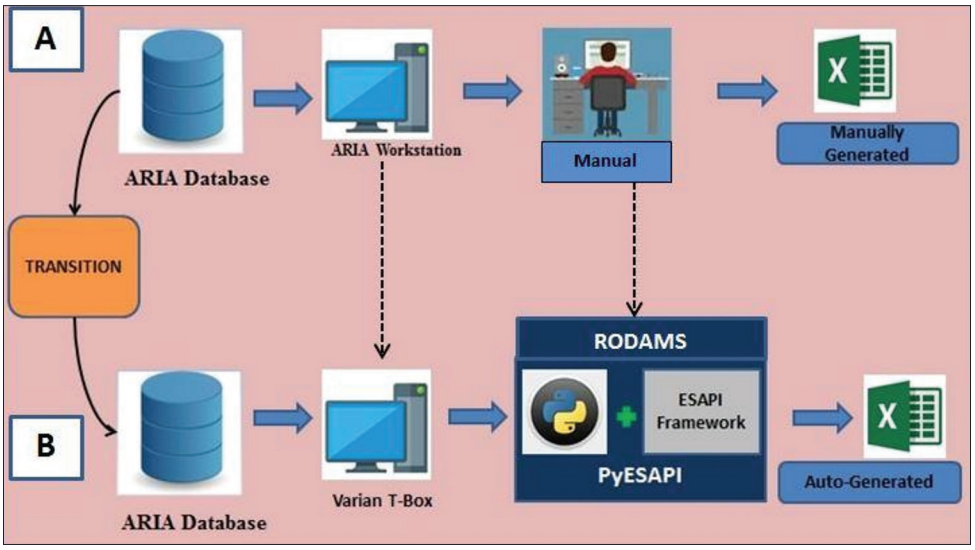

Figure 1: (A) Depiction of the existing manual data extraction process. (B) Representation of the proposed RODAMS-based automatic data extraction process. The dotted line represents the proposed transition from manual data extraction to automated data extraction. RODAMS: Radiation oncology data mining script

Table 2: List of parameters selected for extraction using manual and radiation oncology data mining script techniques in phase-II

Patient and hospital information	External beam treatment unit	Patient's diagnosis details	Treatment course details
Patient's primary ID	Machine unit ID	Clinical description	Course ID
Patient's second ID	Machine model	Diagnosis name	Course intent
Patient's first name	Machine name	Diagnosis code	Collection of plans
Patient's last name	Machine scale name	Diagnosis comment	Treatment start date
Gender	SAD	Dose maximum 3D	Treatment completion date
DOB			
Hospital ID			
Hospital name			
Hospital location			
Primary oncologist ID			
Treatment plan details	**Beam parameters**	**DVH statistics**	
Plan name	Energy name	Maximum dose	
Plan intent	Dose rate	Minimum dose	
Prescribed dose	Meter unit	Median dose	
Number of fraction	Technique	Mean dose	
Approval status	Treatment time	D95%	

SAD: Source to axis distance, DOB: Date of birth, DVH: Dose-volume histogram, 3D: Three dimensional

of features affected the extraction time for both techniques. Specifically, we divided the list of features in Table 2 into groups ranging in size from 10 to 40 features/group. Figure 2 depicts the relationships between various groups.

Methodology

The null hypothesis is that there is no significant difference in the time taken for data extraction between the PyESAPI automated data mining approach and the manual data extraction technique. The alternative hypothesis is that there is a significant difference in the time taken for data extraction between the two techniques. A significance level of 0.05 was used for this study. This study selected a sample size of 427 approved treatment plans. In Phase I, a two-sample *t*-test was used to compare the means of the time required by each technique for data extraction. In Phase II, linear regression was used to examine the relationship between the number of features and extraction time for both techniques.

The data collection process involved randomly assigning patients to each physicist for manual extraction of data, as well as using an automated data mining approach called PyESAPI for data extraction. The time taken for data extraction was recorded for both techniques. The data were analyzed using the chosen statistical test to determine if there is a significant difference in the time taken for data extraction between the two techniques.

Results

The results of this pilot study indicate that the use of the PyESAPI environment and the RODAMS script resulted in a significant improvement in the speed and accuracy of data extraction compared to the traditional manual method.

In Phase I, the average time taken for manual extraction to extract 25 features as shown in Table 1 from the TBox system was 274 ± 19.8 s/plan, whereas RODAMS took an average of only 0.04 ± 0.0005 s/plan. The RODAMS script took 0.28 ± 0.03 min to fetch 25 features from 427 approved plans, in contrast to 1950 ± 140 min taken by manual data extraction for the extraction of 25 features from 427 approved plans. A semi-logarithmic histogram plot comparing the extraction time of manual and RODAMS techniques is presented in Figure 3. The *t*-test revealed that the means of the extraction time were significantly different ($P < 0.001$), with the RODAMS method being 6850 times faster than manual extraction. The accuracy of data extraction through RODAMS was found to be 100% with no missing data, whereas the quality of data extracted manually was inferior due to transcriptional and transpositional errors, incorrect entries, and missing data.

The impact of increasing the number of features on the extraction time for both techniques in Phase II is presented in Table 3. Manual extraction showed a 400% increase in extraction time when the number of features increased from 10 to 40, whereas the RODAMS script exhibited a 34% increase in extraction time. Doubling the number of features extracted manually resulted in an almost 2.5-fold increase in extraction time, whereas the Python script only demonstrated a 1.15-fold increase. A linear regression analysis was conducted to examine the relationship between the number of features and extraction time for both techniques. A scatter plot was constructed with the number of features on the X-axis and extraction time on the Y-axis, as displayed in Figure 4. Both techniques displayed a positive correlation between increasing extraction time and the number of features. The $P < 0.001$ for both techniques indicates that the relationship between the number of features and extraction time is statistically significant.

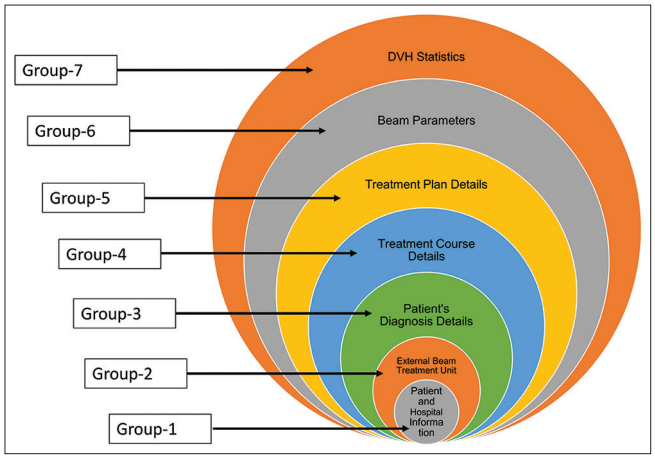

Figure 2: A stacked Venn diagram illustrating the relationships among the various groups of features extracted during Phase II of the study. DVH: Dose-volume histogram

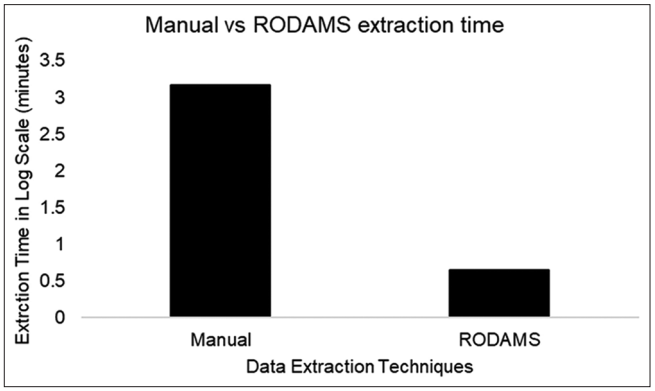

Figure 3: A stacked Venn diagram illustrating the relationships among the various groups of features extracted during Phase II of the study. RODAMS: Radiation oncology data mining script

Table 3: Manual and Radiation Oncology Data Mining Script extraction time versus features statistics for 427 treatment-approved plans

Features	10	15	20	25	30	35	40
Manual extraction time (min)							
Mean	623.9	790.0	1067.5	1470.8	1983.2	2372.2	2533.5
SD	181.8	216.7	206.8	175.0	210.0	179.1	249.4
Maximum	811.3	1024.8	1281.0	1636.8	2135.0	2562.0	2775.5
Minimum	448.4	597.8	868.2	1288.1	1743.6	2206.2	2277.3
RODAMS extraction time (min)							
Mean	0.24	0.25	0.26	0.28	0.27	0.31	0.32
SD	0.01	0.02	0.01	0.01	0.02	0.01	0.01
Maximum	0.26	0.27	0.28	0.29	0.29	0.33	0.33
Minimum	0.23	0.24	0.25	0.26	0.26	0.30	0.30
Ratio of mean extraction time (manual/RODAMS)	2618	3181	4106	5348	7300	7652	7959
Manual normalized time	1.00	1.27	1.71	2.36	3.18	3.80	4.06
RODAMS normalized time	1.00	1.04	1.09	1.15	1.14	1.30	1.34

RODAMS: Radiation oncology data mining script, SD: Standard deviation

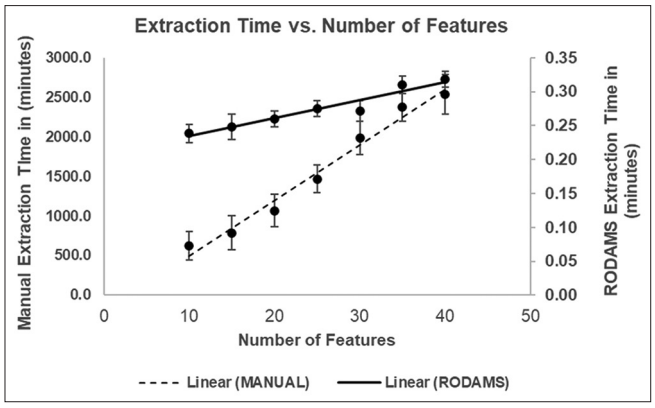

Figure 4: The combo scatter plot graph displays the linear regression of data extraction time in minutes versus number of features for two techniques: Manual and automatic (RODAMS). The graph shows that the RODAMS has a lower slope than the manual technique, indicating that it is faster and more efficient. RODAMS: Radiation oncology data mining script

Discussion

RO is known for its highly quantitative clinical workflow for treatment planning and delivery, making it a suitable field for data analysis and research. However, manual extraction of data from the existing devices and platforms can be time-consuming, prone to errors, and lead to mental burnout for the individuals involved. Programming tools can be applied to efficiently extract data from these devices, as demonstrated in this study using the PyESAPI environment.

In a number of studies, the use of automation in data extraction processes has been demonstrated to be superior to manual extraction methods. According to a study, automation was found to significantly reduce transcription errors and increase the reliability and accuracy of data compared to manual methods.[14] This was attributed to the ability of automated systems to standardize data collection and reduce human-related errors. Another study showed that automation led to a significant reduction in the time required for data extraction and improved the efficiency of the process, while also reducing the stress and workload on the individuals involved.[15] Several studies showed the use of the PyESAPI programming interface API in data extraction, treatment plan analysis, and automatic plan creation.[16,17]

These results are supported by the findings of our own study, which showed that the use of an API, specifically PyESAPI, in the field of RO resulted in more efficient and reliable data extraction compared to manual methods. PyESAPI was able to quickly and accurately extract data from a large clinical database, reducing the time and effort required for data extraction and increasing the accuracy and reliability of the data.

The ARIA system, like other ROIS, stores data in both unstructured and structured ways. Unstructured data, such as pathology reports, is not the focus of this study, but natural language processing can be used to extract information from it. Structured data, on the other hand, can be automatically extracted with the PyESAPI environment.

The PyESAPI environment has limitations, such as being restricted to Varian TBOX in research mode and requiring a data transfer routine and external hard drive for larger clinical databases. Despite these limitations, it offers significant benefits to researchers in RO, allowing for the development of programming scripts for data mining and other applications. However, researchers must have knowledge of the data format and schema, permission to view and extract the data, and software for data mining. The PyESAPI script can sync the ROIS-TPS-API and add strength to clinical utility and research, but further validation from multiple institutions is needed to fully understand the potential of Artificial Intelligence (AI) tools in RO.

Conclusion

The study demonstrated a novel data-mining technique based on the PyESAPI environment and showed its superiority in comparison to manual extraction methods. The PyESAPI script was able to efficiently extract data from the Eclipse planning system, reducing the time and effort required for manual extraction while also increasing the reliability and credibility of the data. The study also highlighted the potential for using the Python programming language and its powerful libraries to extract data in RO. Despite some limitations, the PyESAPI environment offers significant advantages and opportunities for AI learners and young researchers in the field of RO, including the potential for automation in workflow in the RT department. However, further validation from more institutions is needed to fully understand the potential of PyESAPI in the larger context of evolving big data and AI tools in RO.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. McNutt TR, Bowers M, Cheng Z, Han P, Hui X, Moore J, *et al.* Practical data collection and extraction for big data applications in radiotherapy. Med Phys 2018;45:e863-9.
2. Hayman JA, Dekker A, Feng M, Keole SR, McNutt TR, Machtay M, *et al.* Minimum data elements for radiation oncology: An American society for radiation oncology consensus paper. Pract Radiat Oncol 2019;9:395-401.
3. Zapletal E, Bibault JE, Giraud P, Burgun A. Integrating multimodal radiation therapy data into i2b2. Appl Clin Inform 2018;9:377-90.
4. Tagliaferri L, Kovács G, Autorino R, Budrukkar A, Guinot JL, Hildebrand G, *et al.* ENT COBRA (consortium for brachytherapy data analysis): Interdisciplinary standardized data collection system for head and neck patients treated with interventional radiotherapy (brachytherapy). J Contemp Brachytherapy 2016;8:336-43.
5. 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 (R Coll Radiol) 2017;29:814-7.
6. Price S, Golden B, Wasil E, Zhang HH. Data Mining to Aid Beam Angle Selection for Intensity-Modulated Radiation Therapy. ACM BCB

2014 – 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics; 2014. p. 351-9.
7. Shirato H, Le QT, Kobashi K, Prayongrat A, Takao S, Shimizu S, *et al.* Selection of external beam radiotherapy approaches for precise and accurate cancer treatment. J Radiat Res 2018;59:i2-10.
8. Singh N, Kumar S, Bhadauria S. Early detection of cancer using data mining. Int J Appl Math Sci 2016;9:47-52. Available: http://www.ripublication.com. [Last accessed on 2023 Feb 24].
9. Kim KH, Lee S, Shim JB, *et al*. A text-based data mining and toxicity prediction modeling system for a clinical decision support in radiation oncology: A preliminary study. Journal of the Korean Physical Society 2017;7:1231-7. https://doi.org/10.3938/jkps.71.231.
10. Mayo CS, Phillips M, McNutt TR, Palta J, Dekker A, Miller RC, *et al.* Treatment data and technical process challenges for practical big data efforts in radiation oncology. Med Phys 2018;45:e793-810.
11. Roelofs E, Persoon L, Nijsten S, Wiessler W, Dekker A, Lambin P. Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. Radiother Oncol 2013;108:174-9.
12. Osman AF. Radiation Oncology in the Era of Big Data and Machine Learning for Precision Medicine'. Artificial Intelligence - Applications in Medicine and Biology, IntechOpen, 2019. Crossref, doi:10.5772/intechopen.84629.
13. Pyyry J, Wayne Keranen W, editors. A Handbook for Programming in the Varian Oncology Software Ecosystem. The Legrand Orange Book:Varian APIs; 2018.Available from: https://vdocument.in/varian-apis-varian-apis-a-handbook-for-programming-in-the-varian-oncology-software.html?page=1. [Last accessed 2022 Dec 05].
14. Yin AL, Guo WL, Sholle ET, Rajan M, Alshak MN, Choi JJ, *et al.* Comparing automated versus manual data collection for COVID-specific medications from electronic health records. Int J Med Inform 2022;157:104622.
15. Thorat C, Bhat A, Sawant P, Bartakke I, Shirsath S. A detailed review on text extraction using optical character recognition. Lect Notes Netw Syst 2022;314:719-28.
16. Anchineyan P, Amalraj J, Krishnan BT, Ananthalakshmi MC, Jayaraman P, Krishnasamy R. Assessment of knowledge-based planning model in combination with multi-criteria optimization in head-and-neck cancers. J Med Phys 2022;47:119-25.
17. van Marlen P, Verbakel WF, Slotman BJ, Dahele M. Single-fraction 34 Gy lung stereotactic body radiation therapy using proton transmission beams: FLASH-dose calculations and the influence of different dose-rate methods and dose/dose-rate thresholds. Adv Radiat Oncol 2022;7:100954.