ESTRO 2023
From Innovation to Action
12-16 May 2023 | Vienna, Austria

ESTRO 2023 Physics Track - Big Data, Big Headache Session

Tips for mitigating and avoiding common legacy-data headaches

Legacy treatment data are a necessary headache for many radiotherapy researchers. Whether investigating the link between radiotherapy exposures and late effects with long latency periods, looking to bolster numbers for statistical power when studying rare diseases and/or effects, or seeking to improve data variety for advanced analytical techniques like data mining and deep learning, legacy data including images, structure contours, and dose distributions associated with historical, retired treatment planning systems (TPSs), may represent a gold mine of sorely needed data. Despite their value, working with legacy data is likely to involve several obstacles that must be overcome before their utility can be realised.

We found it necessary to leverage treatment data created as many as 25 years ago for all of the aforementioned reasons. We initially identified 464 patients treated as far back as 1996 who were eligible for our study. Our journey to recover treatment planning data for these patients led us to face three broad types of headaches, namely those involving the complete loss of data, others involving the partial loss of data, and more still involving distorted data. Incidents of complete data loss resulted from catastrophic occurrences like major server crashes, which could only have been mitigated by prior prospective measures. Those affected data were completely unrecoverable and ultimately led to cohort shrinkage. Partial data loss often resulted from data-transfer anomalies, such as differences between the native format, in which data were stored in the legacy treatment planning system, and modern DICOM format in which data were exported. This resulted in missing information in the DICOM header upon export, including image dates and incomplete renderings of structure contours. Luckily, missing data components could often be recovered via ancillary measures. In the case of missing image dates, we most often recovered the true image date from the Picture Archiving and Communication System (PACS) record of the same image and, when that was unavailable, we could reasonably approximate it using the date on which radiotherapy started. Because these additional measures were often needed for all patients, we found it useful to use a small test cohort to practice the export and data analysis pipelines and verify that all clinical variables were retrieved, as expected. Once all missing data components were identified in the test cohort, we refined our pipelines to involve parallel measures to collect the needed data alongside export for the remainder of the cohort.

The final type of headache, related to distorted data, seemed to be the most abundant and encompassed wide-ranging manifestations requiring equally diverse mitigation measures. Some common forms of data distortion include erroneous dose accumulations and spatial transformations. Although the root cause of erroneous dose accumulations was unclear, we could often rectify them by retrieving the archived files from each phase of treatment in the retired TPS, exporting them individually, and accumulating the course outside of the legacy TPS. Spatial-transformation distortions represented data for which the treatment planning image, structure set, and dose distribution were spatially misaligned after export. Luckily, the transformations needed to realign the data matrices were reproducible across patients, allowing for a class solution that could be applied whenever spatial transformation distortions were recognised. Generally speaking, data distortions arose on a case-by-case basis, were difficult to anticipate, and could be remedied using retrospective measures. Therefore, we found it useful to screen data for possible distortion anomalies using simple dose-volume tests. More specifically, we stratified patients by treatment protocol and considered which dosimetric characteristics we expected to be shared within each stratum. We then plotted histograms of dose-volume measures that would probe those expected characteristics such that outliers would highlight cases that likely needed additional attention.

In the end, we were able to recover complete, validated radiotherapy data for 268 patients. Looking forward, we have implemented a procedure and infrastructure to collect complete volumetric radiotherapy treatment data that are duly validated, organised, and backed up for future researchers to access without the need to reach into legacy systems or decipher complex treatment courses. The new infrastructure can accommodate legacy data, such as those recovered here, as well as newly planned treatment data to create a master database specifically designed to avoid future legacy data headaches.

Lydia J Wilson
Medical Physicist
St. Jude Children's Research Hospital, Memphis, TN
Thomas Jefferson University, Philadelphia, PA
USA