advances
in radiation oncology

Scientific Article

Big Data Readiness in Radiation Oncology: An Efficient Approach for Relabeling Radiation Therapy Structures With Their TG-263 Standard Name in Real-World Data Sets

Thilo Schuler MD [a,b,]*, John Kipritidis PhD [a], Thomas Eade MD [a,c],
George Hruby MD [a,c], Andrew Kneebone MD [a,c], Mario Perez [a],
Kylie Grimberg [a], Kylie Richardson [a], Sally Evill [a], Brooke Evans [a],
Blanca Gallego PhD [b]

[a]*Department of Radiation Oncology, Northern Sydney Cancer Centre, Royal North Shore Hospital, Sydney, Australia;*
[b]*Australian Institute of Health Innovation, Macquarie University, Sydney, Australia; and* [c]*Northern Clinical School,
University of Sydney, Sydney, Australia*

Abstract

Purpose: To prepare for big data analyses on radiation therapy data, we developed *Stature,* a tool-supported approach for standardization of structure names in existing radiation therapy plans. We applied the widely endorsed nomenclature standard TG-263 as the mapping target and quantified the structure name inconsistency in 2 real-world data sets.

Methods and Materials: The clinically relevant structures in the radiation therapy plans were identified by reference to randomized controlled trials. The Stature approach was used by clinicians to identify the synonyms for each relevant structure, which was then mapped to the corresponding TG-263 name. We applied Stature to standardize the structure names for 654 patients with prostate cancer (PCa) and 224 patients with head and neck squamous cell carcinoma (HNSCC) who received curative radiation therapy at our institution between 2007 and 2017. The accuracy of the Stature process was manually validated in a random sample from each cohort. For the HNSCC cohort we measured the resource requirements for Stature, and for the PCa cohort we demonstrated its impact on an example clinical analytics scenario.

Results: All but 1 synonym group ("Hydrogel") was mapped to the corresponding TG-263 name, resulting in a TG-263 relabel rate of 99% (8837 of 8925 structures). For the PCa cohort, Stature matched a total of 5969 structures. Of these, 5682 structures were exact matches (ie, following local naming convention), 284 were matched via a synonym, and 3 required manual matching. This original radiation therapy structure names therefore had a naming inconsistency rate of 4.81%. For the HNSCC cohort, Stature mapped a total of 2956 structures (2638 exact, 304 synonym, 14 manual; 10.76% inconsistency rate) and required 7.5 clinician hours. The clinician hours required

* Corresponding author. Department of Radiation Oncology, Northern Sydney Cancer Centre, Reserve Rd, Sydney 2065, Australia.
  E-mail address: thilo.schuler@health.nsw.gov.au (T. Schuler).

were one-fifth of those that would be required for manual relabeling. The accuracy of Stature was 99.97% (PCa) and 99.61% (HNSCC).

Conclusions: The Stature approach was highly accurate and had significant resource efficiencies compared with manual curation.

Introduction

Modern radiation therapy (RT) uses complex techniques such as intensity modulation and image guidance to achieve precise and highly conformal dose distributions. The basis for these techniques are 3-dimensional computer models of target and organ-at-risk (OAR) structures. Consistent naming of these RT plan structures (henceforth referred to as *structures*) improves patient safety through better communication[1] and through the ability to perform automated quality assurance.[2] It also streamlines the implementation and use of RT planning decision-support tools such as knowledge-based planning.[3,4] Consistent structure naming is critical for research on dosimetric data, both in prospective clinical trials,[5] or when using observational big data methods on routinely collected data at the institutional and interinstitutional level.[6,7] Although the dosimetric data are automatically and consistently captured in high quality in the treatment planning system (TPS) because of its direct interface with the treatment machines (linear accelerators), the names of the RT plan structures may vary widely because they are manually entered.

Structure name inconsistency is challenging to resolve in an automated way because of the difficulty in distinguishing between typographic name variations and genuine semantic differences. For example, adding the initials of a trainee physician or an abbreviation of a planning or delivery technique (eg, RP for "Rapid Plan" or RA for "Rapid Arc") to a structure signals that this structure should be omitted for dosimetric analysis. On the other hand, the initials of a consulting physician or an abbreviation such as PC (for "postchemotherapy") may be used for clinically meaningful structures.

Lack of structure name consistency has been identified as a problem in clinical trials despite strict protocols. This has led the major RT trials organization NRG/Radiation Therapy Oncology Group to employ tools to review and enforce adherence to their structure name library.[5,8] Without the enforceable rules of clinical trials, structure name inconsistency is greater in routine care,[9] although we are not aware of studies that quantify this problem. Reasons for this include varying clinician preferences, which—particularly in large institutions—can make consensus challenging. If a local policy exists, there is a

risk of it becoming outdated when new treatments are adopted. In busy routine practice, time and know-how are often limited to establish systems that optimally support adherence to local policies. To reduce inconsistency from the outset requires complex, multiyear change management projects that involve all stakeholders[6,10] and establish the right culture to maximize consistency.[7] Structure name inconsistencies are compounded when seeking to conduct analysis of pooled RT plan data from multiple institutions.

In January 2018 task group number 263, convened by the American Association of Physicists in Medicine, released its final report on a nomenclature system for target and OAR volumes and dose-volume histogram (DVH) nomenclature[11] (henceforth referred to as TG-263). Its development involved a comprehensive multistakeholder process that included representatives of all major organizations in the field of radiation oncology. When implemented by institutions, it will play an important part in enabling data pooling on a large scale.

Applying strict naming protocols that are linked to global nomenclatures, kept up to date with current clinical practice, and supported by technological and workflow practices (eg, automatic quality assurance mechanisms) will be the best way of dealing with the structure name inconsistency problem going forward. However, these changes are only being gradually implemented and will not solve the problem of inconsistent historical RT plan data. The typical solution for correcting historical data is to manually relabel structures, which is very resource intensive and therefore prohibitive for larger data sets.[9] To our knowledge there are no published reports on more scalable approaches specifically designed to address this problem. To meet this need, we have developed Stature, a tool-supported approach for scalable standardization of structure names in existing RT plans. In this paper we

1. describe the Stature approach to retrospectively standardize structure names;
2. demonstrate the utility of Stature in quantifying and overcoming the problem of structure name inconsistency in prostate cancer (PCa) and head and neck cancer data sets by mapping to the TG-263 standard names; and
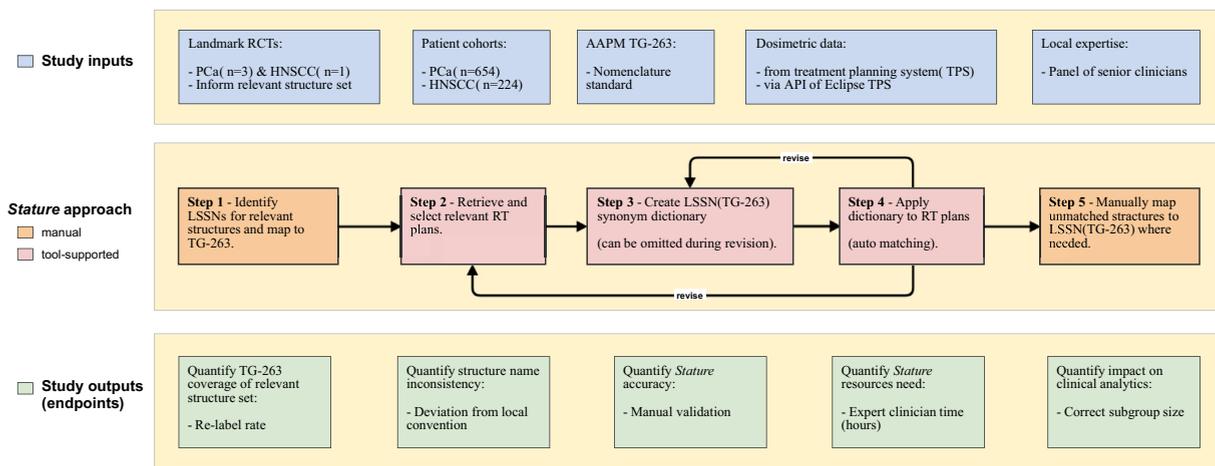3. illustrate the impact on a clinical analytics scenario.

Figure 1   Visual schematic of the Stature approach with specific inputs and outputs for this study. *Abbreviations:* API = application programming interface; HNSCC = mucosal head and neck squamous cell carcinoma; LSSN = local standard structure name; PCa = prostate cancer; RCT = randomized controlled trial; RT = radiation therapy.

Methods and Materials

Data

After institutional review board approval (NSLHD HREC reference: LNR/15/HAWKE/355), we included patients with PCa and mucosal head and neck squamous cell carcinoma (HNSCC) who received conventionally fractionated (1.8-2 Gy per daily fraction), curative RT at our institution between 2007 and 2016 (PCa) and 2007 and 2017 (HNSCC). All patients received an intensity modulated RT technique. PCa patients were treated with definitive RT with (n = 115) and without (n = 338) pelvic nodal irradiation and with postprostatectomy RT with (n = 34) and without (n = 167) pelvic nodal irradiation. HNSCC patients who received postoperative RT were excluded. In total 654 PCa patients and 224 HNSCC patients (henceforth referred to as the PCa cohort and the HNSCC cohort) met these criteria.

The Stature approach

The central idea behind the Stature approach for retrospective standardization of RT structure names was to capture and apply senior clinicians' knowledge of structure names with the aid of a software tool. We opted for this design choice because an interdisciplinary panel of local clinicians is best suited to decode the meaning of RT structure name variants, which are specific to the local and temporal context. The software tool was developed to be time efficient by enabling scalable structure name review. Software development was performed in-house using the C# programming language (Microsoft Corp, Redmond, WA) and the scripting application programming interface of our TPS (Eclipse V13.6, Varian

Medical Systems, Palo Alto, CA). The Stature approach consists of 5 steps, as displayed in Figure 1. Steps 2 to 4 are supported by the Stature tool. The tool's source code has been made available under the Apache 2.0 Open Source license in the following public github repository: https://github.com/tschuler/StatureTool.

Step 1: Identify relevant structures requiring TG-263 labeling

The Stature approach started with the identification of a set of clinically relevant structures to be relabeled with their TG-263 name. To focus our work on the structures that are relevant to clinical research, 2 radiation oncology doctors reviewed the protocols of large contemporary randomized controlled trials (RCTs) and identified the structure names that they contained. This comprised 15 structures from 3 PCa RCTs[12–14] (we combined the structures from the 3 RCTs to cover our full treatment spectrum) and 14 structures from an HNSCC RCT.[15] For both cohorts, we then expanded the set by 3 to 4 structures that were important to our local practice. We manually mapped the resulting 36 structures to their TG-263 name.

In an optional substep we also compared each of these 36 structures to our local naming convention and determined the corresponding local standard structure name (LSSN) that had been established at our institution before the availability of TG-263. We performed this intermediary LSSN mapping step to enable the quantification of past naming inconsistency. To allow nuanced reporting on the past naming consistency, the LSSNs were classified according to their importance for routine care as "essential" (near 100% coverage is expected), "optional" (coverage depends on clinical scenario), and "derived" (volume is not routinely contoured as can be created from other structures).

Step 2: Retrieve and select relevant RT plans

The Stature tool retrieved all RT plan data for the included patients, which were identified through the clinical information systems used for prostate and head and neck cancer at our institution. These RT plan data were then filtered, based on attributes such as "treated" status or fractionation details, to select the relevant plan for each patient. For instance, we excluded RT plans that were never used to deliver radiation to a patient.

Step 3: Create TG-263/LSSN synonym dictionary

We established an interdisciplinary panel of senior clinicians consisting of 1 radiation oncologist, 1 medical physicist, and 1 radiation therapist (dosimetrist). Each senior expert had longstanding local experience in treating PCa and HNSCC. The radiation oncology trainee who implemented the Stature tool complemented the panel as tool expert.

In an iterative process the expert panel identified structure synonyms, which were used to create a "lookup dictionary" that associated these synonyms with the corresponding LSSN, which in turn had been mapped to TG-263. If the sole purpose is to map to TG-263, this intermediary LSSN mapping could be omitted. In that case the dictionary would associate the synonyms for a structure, including the LSSN, directly with its TG-263 name.

Interactive dictionary creation was facilitated by a frequency analysis feature in Stature, which produced a ranked list of structure names across all plans selected in step 2. The list of structure names was then filtered by name (using regular expressions), dose, and volume attributes to narrow the list to synonym candidates for a particular LSSN. Drawing on their knowledge of local practice, the clinical experts would then iteratively remove irrelevant or redundant names to arrive at a set of synonyms for each LSSN. In addition to using certain dose and volume metrics during interactive dictionary development by the experts, these attributes could also be included into the dictionary as cutoff rules, allowing the tool to make metrics-based synonym grouping decisions during step 4. From a tool implementation point of view we achieved this by including an existing JavaScript interpreter (https://github.com/sebastienros/jint) that, at runtime, evaluates expert-defined rules based on dose or volume cutoffs (such as D95, Dmean, or volume).

Step 4: Apply TG-263/LSSN synonym dictionary to RT plans

The lookup dictionary was applied to all relevant plans selected in step 2, and the match results were visualized as a heatmap with structures classified as having an exact LSSN match (green cells), having a synonym LSSN match (orange cells), requiring manual intervention (yellow cells), or having no LSSN match (red cells). Unexpected heatmap patterns (ie, plans containing clusters of unmatched structures) were inspected by drilling down to the structure-specific dose and volume data for that plan. Where mismatches or omissions were identified, either the LSSN synonym dictionary (step 3) or the RT plan selection (step 2) were revised. Figure 2 illustrates a screenshot of the tool's heatmap visualization and the interactive drill-down area, which allows the LSSN matches to be viewed in the context of other structures.
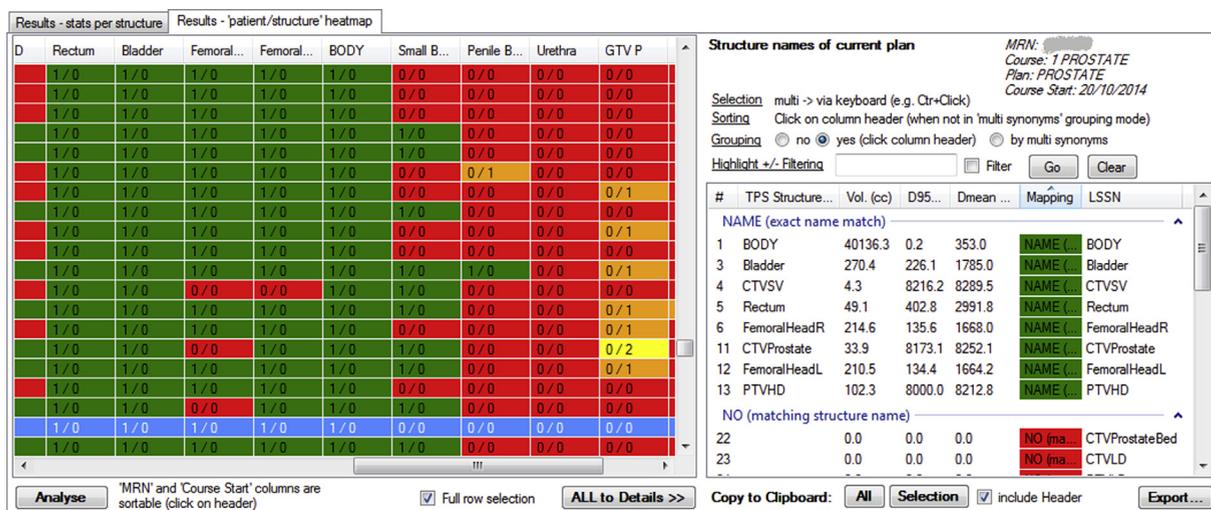

Figure 2   Screenshot from Stature tool. The left panel displays the heatmap visualization where each row refers to a specific plan and each column refers to a given LSSN. Cells are colored green, orange, yellow, or red for an exact LSSN match, synonym match, manual match, or nonmatch, respectively. The right panel displays the drill down area providing structure-specific dose and volume data for each plan. *Abbreviations:* LSSN = local standard structure name; TPS = treatment planning system.

For cases in which *Stature* could not resolve the non-matched structures, the plan was reviewed in the TPS.

Step 5: Manual matching intervention

Manual matching was required in situations in which the tool could not automatically match an RT plan to a particular LSSN because the plan included more than 1 synonym for this LSSN.

Metrics and validation

Using descriptive statistics, we compiled metrics on the process and outcomes of applying Stature to the PCa and HNSCC cohorts as follows:

1. Percentage of relevant structures mapped to TG-263. A measure of TG-263's coverage of the identified relevant structures across the PCa and HNSCC use cases.

2. Structure name inconsistency. We defined structure name inconsistency as the percentage of structures that did not follow the local naming convention. Equation 1 shows the calculation performed to derive the inconsistency rate. This inconsistency rate was calculated (across all plans) for (1) all LSSNs combined, for (2) the subset of essential and optional LSSNs, and (3) on a per-LSSN basis.

HNSCC cohort every structure in the sampled plans was checked in the TPS. For the PCa cohort the TPS was only consulted if the match or nonmatch appeared unusual. A radiation oncology trainee arbitrated to resolve any discordance.

4. Resource requirements. For the HNSCC cohort we measured the resource requirements for Stature in terms of people and time taken to perform steps 2 to 4.

5. Impact on clinical analytics. To demonstrate the utility of Stature for use in clinical analytics, we used our clinical data warehouse (CDW), which was developed at our institution using a MySQL Database (Oracle, Redwood City, CA). The CDW can mirror the structure DVH data from the TPS and was provided the same set of PCa plans selected in step 2 of the Stature approach. The Stature-derived synonym dictionaries were applied to these plans in the CDW. To identify Stature's impact on an example analysis of population-level DVH data, we compared the output of the CDW queries with and without the use of the lookup dictionaries: The population DVH curves were created with the statistical package R and one of its DVH libraries.[16,17]

$$\text{Structure name inconsistency}(\%) = \frac{\text{Number of LSSN matches (synonym + manual)}}{\text{Number of all LSSN matches (exact + synonym + manual)}} \times 100 \qquad (1)$$

For the HNSCC cohort we additionally investigated how the structure name inconsistency rate has evolved in 3 different periods (2007-2010, 2011-2013, and 2014-2017).

3. Accuracy of Stature. The accuracy of the tool-based intervention was validated based on a manual review of 2 randomly selected subsets of PCa and HNSCC plans. The subsets comprised 13.5% of PCa plans (n = 88) and 13.4% of HNSCC plans (n = 30). Two radiation therapists who had not been involved in the application of Stature independently identified incorrect or missing LSSN matches. For the

Results

Table 1 summarizes our key results.

Percentage of structures relabeled with their TG-263 name

All LSSNs except "Hydrogel" could be mapped to a TG-263 nomenclature–conformant label. There were 88 matches of "Hydrogel" or its synonyms out of 8925

Table 1   Summary of results

	TG-263 coverage (% LSSNs mapped to TG-263)	Structure name inconsistency	Stature accuracy	Stature resource requirements	Impact on clinical analytics
Prostate cancer cohort	99.01% (8837 of 8925)	4.81%	99.97%	Not measured	Cohort size changes: 77% and 26%
Head and neck cancer cohort		10.76%	99.61%	7.5 clinician hours	Not measured

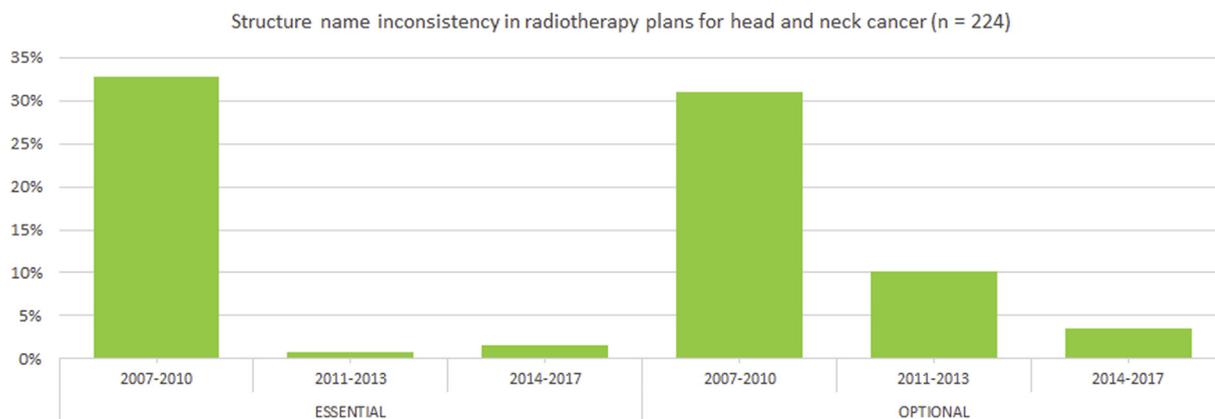*Abbreviation:* LSSN = local standard structure name.

Figure 3 Evolution of structure name inconsistency in head and neck cancer patients stratified by LSSN groups "essential" and "optional." *Abbreviation:* LSSN = local standard structure name.

LSSN matches across both cohorts. Thus 99.0% of relevant structures could be mapped to TG-263.

The LSSNs identified from step 1 in the section The Stature approach, their classification (as essential, optional, or derived), and mappings to TG-263 names are available in Appendix E1 (available online at https://doi.org/10.1016/j.adro.2018.09.013).

Structure name inconsistency

Prostate cancer cohort

The Stature tool performed 5969 LSSN matches in the PCa cohort, of which 5682 structures could be matched exactly, 284 were matched via a synonym, and 3 required manual matching. Thus the overall structure relabeling or inconsistency rate was 4.81%. The structure name inconsistency rate was 1.67% and 47.67% when considering only essential and optional LSSNs, respectively. Despite the low inconsistency rates, there were 3 notable LSSNs with a significantly larger percentage of inconsistently named structures: "GTV P" at 92.5%, "GTV N" at 61.9%, and "Urethra" at 18.9%. These 3 inconsistent LSSNs constituted half the number of LSSNs in the optional group and were the reason for the higher inconsistency in this group. Across all PCa LSSNs the mean number of synonyms was 5.0 (range, 1-12). The structure name inconsistency rates and underlying absolute numbers for each LSSN can be found in Appendix E2 (available online at https://doi.org/10.1016/j.adro.2018.09.013).

Head and neck cancer cohort

The Stature tool performed 2956 LSSN matches in the HNSCC cohort, of which 2683 structures could be matched exactly, 304 were matched via a synonym, and 14 required manual matching. Thus the overall structure relabeling or inconsistency rate was 10.76%. The structure name inconsistency rate was 10.05% and 11.77% when considering only essential and optional

LSSNs, respectively. At the LSSN level the 3 LSSNs with the highest inconsistency were "GTV PREC PET_*P*" at 53.9%, "GTV PREC PET_N" at 50.0%, and "Brainstem" at 42.2%. Across all HNSCC LSSNs the mean number of synonyms was 9.6 (range, 3-14). The structure name inconsistency rates and underlying absolute numbers for each LSSN can be found in Appendix E2 (available online at https://doi.org/10.1016/j.adro.2018.09.013).

Figure 3 shows the change of structure name inconsistency of HNSCC LSSNs classified as essential and optional. The inconsistency was highest and similar between the 2 groups at around 30% in the first interval. In the essential group the inconsistency was minimal in the subsequent intervals. In the optional group, which includes the integrated positron emission tomography and computed tomography (PET/CT)–based structures, the inconsistency was reduced more slowly.

Accuracy of Stature

A random sample of 88 RT plans from the PCa cohort (13.5% of total PCa cohort) was manually checked. Across both independent reviewers there were 3 incorrect LSSN matches compared with the Stature matching results. After arbitration review, only 1 match was truly incorrect out of a total of 3457 sampled structures. Hence the accuracy of the tool-based matching intervention for the PCa cohort was 99.97%.

A random sample of 30 RT plans from the HNSCC cohort (13.4% of total HNSCC cohort) was manually checked. Across both independent reviewers there were 5 incorrect LSSN matches compared with the Stature matching results. Arbitration review confirmed 5 truly incorrect matches out of a total of 1294 sampled structures. Hence the accuracy of the tool-based matching intervention for the HNSCC cohort was 99.61%.
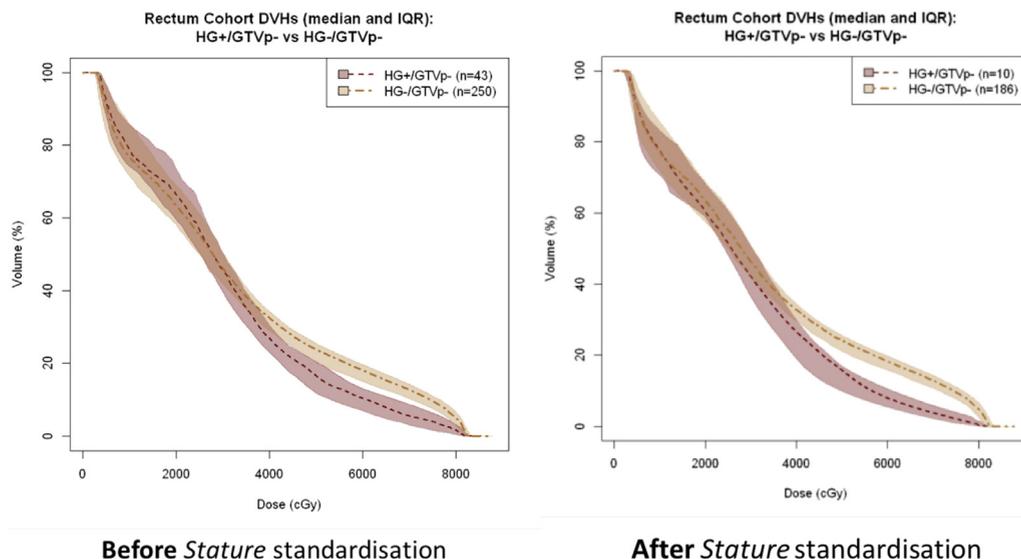
Figure 4   Comparing rectum cohort dose-volume histograms (DVHs) based on hydrogel (HG) and GTV boost (GTVp) status. *Abbreviation:* IQR = interquartile range.

Resource requirements

The resource requirement for HNSCC-cohort Stature approach was 7.5 clinician hours, which was 3.5 hours for the tool expert to prepare the data for the expert panel meeting, and for postmeeting revision, 1 hour each for the 4 experts in the panel meeting.

As a point of comparison, the process of manually checking the HNSCC sample, which included review of the structures in the TPS, took the 2 trained radiotherapists 305 and 325 minutes, respectively (mean 315 minutes). Extrapolating this figure to the full cohort of 224 patients would result in 2352 minutes or 39.2 hours. This is a conservative proxy for the resource requirements of manual relabeling because it does not account for the additional time needed to unapprove and reapprove historical plans.

Therefore Stature performed the structure standardization task for the whole HNSCC cohort with 19% of the clinician resources needed for manual relabeling (7.5 vs 39.2 hours).

Impact on clinical analytics

In an example clinical analytics scenario that compared population DVH data, we identified those PCa patients receiving definitive RT without pelvic nodal irradiation and without an integrated gross tumor volume (GTV) boost (GTVp−) based on the absence of the LSSN "GTV P." We then separated the DVH data for patients who had implantation of a hydrogel spacer (HG+) or no hydrogel spacer (HG−) based on the presence or absence of the LSSN "hydrogel."

When structuring the CDW queries based on an exact LSSN match only (ie, no synonym dictionary), the HG+/GTVp− and HG−/GTVp− groups consisted of 43 and 250

patients, respectively. When the LSSN synonym dictionary was included, these numbers decreased to 10 patients (77% change) and 186 patients (26%), respectively. Without the LSSN normalization the cohort without GTV boost appeared bigger because the patients with "GTV P" synonyms were erroneously included. In this case this misclassification would have reduced the gap between the 2 population medians seen in Figure 4.

Discussion

Retrospectively improving data quality, also known as curation, is typically performed manually, consuming significant time for larger data sets. Stature, an approach to standardize existing RT structures with their TG-263 name using a software tool to capture and apply clinician knowledge, was successfully applied in PCa and HNSCC patients who received curative-intent RT at an academic center in Australia between 2007 and 2017.

Selection of clinically relevant structures

Apart from the 7 additional structures (all classified as optional) that were added to reflect contemporary techniques at our center, the selection of the structures was based on landmark RCTs. All mandatory structures in the RCT protocols were also deemed essential, except for 2 structures (penile bulb, brainstem) for which contouring depends on clinical circumstances in our practice.

TG-263 mappings

Almost all relevant structures (99.0%) could be relabeled with their TG-263 name. The only structure without

a TG-263 name was "Hydrogel." This artificial structure represents a hydrogel-based spacer injected between rectum and prostate to increase separation between these organs. Although neither a true target nor an OAR, we included it because it is relevant for clinical analyses. We have contacted the TG-263 chair, and "Hydrogel" will be considered by the committee for inclusion in future.

Routine care structure name consistency

The overall naming consistency across all examined PCa structures was high. However, the few structures with low consistency can have a negative effect on clinical analytics, as demonstrated in the rectum cohort DVH use case. In line with expectations, the structure naming consistency in the HNSCC setting, which has inherently more complex RT plans, was lower. Other factors for this difference are a strong academic focus on PCa by a small subspecialty team of 3 radiation oncologists in our institution. This interest includes dosimetric effects,[18-20] which was an early driver for consistent naming. Other areas within the department have often followed this lead and implemented similar measures. In the case of head and neck cancer RT, cultural and systems changes, such as naming templates, have led to improvements in structure name consistency, as shown in Figure 3.

The TG-263 report recognizes the additional complexities in the naming of RT targets compared with OARs. Our results support this observation, finding the highest name variability for target structures, particularly in areas where the clinical practice had recently changed. Examples of recent changes in our practice are dose painting with higher doses to dominant intraprostatic tumor nodules and involved lymph nodes in PCa because of the introduction of gallium-68 prostate-specific membrane antigen PET/CT (PSMA-PET/CT) in late 2014 or HNSCC tumor delineation aided by fluorine-18 fluorodeoxyglucose PET/CT (FDG-PET/CT). To some degree this is inevitable, but clinician awareness and agile change management, including adaptation of systems, would reduce the resulting structure name inconsistency. Triggered by the results from this study, we intend to introduce tumor- and RT type–specific structure name conformance checks at the time of RT plan approval. This will be facilitated through a custom TPS plugin.

Resource requirements

The matching accuracy of the Stature approach was very high because of interdisciplinary expert involvement and sufficient iterative refinements. The time required from senior domain experts remained less than the expected range of 2 hours per person. This means it will be feasible to apply the Stature approach, over time, to the historical RT plans from all treatment sites in our institution. The tool expert had to invest more time than the senior domain experts, and there is an overall tradeoff between time spent on iterations and mapping accuracy: More time spent on iterations leads to a higher mapping accuracy, although with diminishing returns per hour. This became particularly apparent in the head and neck cancer scenario, and a goal-informed time limit was proposed for future efforts. For a treatment site with low patient numbers, the goal might be to invest more time into the curation effort to maximize patient numbers for optimal statistical analysis. However, for high-volume treatment sites, it might be acceptable not to extract data for every patient. Thus some incorrect nonmappings of potentially mappable structures are tolerated for the sake of speeding up the curation effort. Extrapolating the time for manual validation of the HNSCC cohort to the full cohort provides a conservative estimate of a manual curation effort (39.2 clinician hours). Comparison to the Stature approach (7.5 clinician hours) indicates the benefit, which would increase when curating bigger cohorts.

Stature design choices

It is important to point out that there are other techniques in the medical literature to find synonyms. We considered applying natural language processing methods to increase the level of automation when resolving ambiguity in identifying structure name synonyms.[21] In particular we considered string similarity algorithms and ontology-based similarity using the concept synonyms in the Unified Medical Language System metathesaurus.[22] These methods have been described by Soğancõoğlu et al.[23] The decision not to use these natural language processing methods and to rely on knowledge from local domain experts instead was based on 2 observations:

1. Analysis of RT structure names found highly idiosyncratic patterns not reflected well in Unified Medical Language System. A typical RT plan contains several structures with the same name stem, only distinguished by short qualifiers (typically 1-3 letters) that could change the meaning significantly. This made automation via string similarity less promising. Although regular expression-based identification of synonym candidates during Stature step 3 worked well, this task could potentially be enhanced by a string similarity-driven feature in Stature.
2. Because the scope in Stature was local RT plans, involving local clinical experts was an option as long as the time requirement was minimized.

Other than efficiency, another goal of the Stature tool was to make the curation approach consistent and reproducible when rolling it out across different tumor sites. Creating the dictionary is the crucial aspect of the Stature approach that turns implicit domain knowledge into rules that can be applied in a scalable way. Initially dictionary creation was only based on structure names. During the iterative Stature development using the PCa cohort data, we realized the benefit of augmenting the frequency analysis (step 3) and dictionary application (step 4) by volume and dose attributes, and we subsequently added these features. For step 3 this was achieved by displaying dose-volume information about the currently selected RT plan in a sortable view (see Fig 2, right panel). For step 4 this involved clinical experts defining dose-volume based rules in the dictionary, which the Stature tool interpreted to make decisions regarding classifying an ambiguous structure name. As an example, in our definitive RT plans for PCa we did not commonly specify in the structure name where a GTV was located (primary tumor vs lymph nodes) because it was obvious when looking at the plan. As a result of modern imaging techniques, it is now quite common to have several GTVs. Because our doses to these differ, a dose-based cutoff was used to have the Stature tool retrospectively classify into "GTV N" and "GTV P".

Related work

The focus on retrospective relabeling of existing structures distinguishes this study from several published efforts that aimed at improving structure name consistency prospectively. Mayo et al.[6] described how they successfully standardized structure naming and RT prescription in a 4-year practice change project in their institution. The same group also led the TG-263 effort, which sought input from the multidisciplinary radiation oncology community to develop a standardized nomenclature for the field.[11] One of TG-263s strengths is that it does not purely list names but rather provides a generic schema that allows the addition of structures in the future.

A global standard for structure names is indispensable for interinstitutional data exchange, and our study provides one of the first accounts of mapping to TG-263, which has positioned itself strongly for this task. If TG-263 is adopted widely and its implementation is coupled with the necessary change in culture,[24] this will address the intra- and interinstitutional structure naming inconsistency issue going forward. Our work complements these approaches by addressing naming inconsistencies retrospectively.

Apart from retrospectively mapping historical structure names to TG-263 to aid data exchange, we have decided to adopt TG-263 for any new naming decisions in our clinic. We also aim to gradually change our practice toward using the TG-263 names instead of the locally established structure name conventions. The fact that many of our currently used structure names are already similar to the TG-263 name will help with this.

To date, we are only aware of one other tool-based curation effort in radiation oncology.[25] This evolving project targets data from multiple institutions for pooling into a shared registry. As such, that project dealt with a much wider variety of data types, and local expert knowledge could not be relied on for data integration and deduplication. Instead, various statistical machine learning methods are used to match incoming data with the existing fields of the registry.

Limitations

There are limitations in this work. Stature has only been validated in 2 patient cohorts from 1 relatively small (3 linear accelerators) radiation oncology department with a strong academic focus. It remains to be determined whether the results are generalizable to other settings. The current tool only supports 1 TPS (Eclipse). However, we believe the Stature approach itself is generic and TPS agnostic. Thus a similar tool could be implemented for different TPS. To increase replicability, we have provided the source code of our tool (see link in The Stature Approach). Although Stature reduces the resource need compared with manual curation, it does not eliminate it, and not every department will have the local expert resources required for Stature's semiautomated, expert knowledge–driven approach. The need for local expert knowledge also prohibits Stature's use when pooling data from different institutions, unless each contributing institution uses Stature locally before pooling (bottom-up curation).

Directions

We intend to apply Stature to the clinical plans of all RT courses that were delivered in our institution in the last 10 years. Having a large CDW with standardized structure names will enable real-time cohort studies[26] that can provide decision support in the treatment planning stage, which is an additional benefit to using the CDW for formal clinical research. We also plan to explore whether supervised machine learning using dose and volume information can enhance the expert-based approach or potentially completely replace it.

Conclusions

We developed and validated a retrospective RT structure name standardization approach called Stature. This work indicates that Stature is feasible, saves time compared with manual curation, and is beneficial for

clinical analytics. It also quantifies the extent of structure name inconsistency in the routine care setting.

Acknowledgments

The authors thank Mary-Anne Brancato for assisting in the preparation of this report.

Supplementary data

Supplementary material for this article can be found at https://doi.org/10.1016/j.adro.2018.09.013.

References

1. Potters L, Ford E, Evans S, Pawlicki T, Mutic S. A systems approach using big data to improve safety and quality in radiation oncology. *Int J Radiat Oncol.* 2016;95:885-889.
2. Olsen LA, Robinson CG, He GR, et al. Automated radiation therapy treatment plan workflow using a commercial application programming interface. *Pract Radiat Oncol.* 2014;4:358-367.
3. Appenzoller LM, Michalski JM, Thorstad WL, Mutic S, Moore KL. Predicting dose-volume histograms for organs-at-risk in IMRT planning. *Med Phys.* 2012;39:7446-7461.
4. Tol JP, Dahele M, Delaney AR, Slotman BJ, Verbakel WFAR. Can knowledge-based DVH predictions be used for automated, individualized quality assurance of radiotherapy treatment plans? *Radiat Oncol Lond Engl.* 2015;10.
5. Giaddui T, Yu J, Manfredi D, et al. Structures validation profiles in transmission of imaging and data (TRIAD) for automated NCTN clinical trial digital data quality assurance. *Pract Radiat Oncol.* 2016;6:331-333.
6. Mayo CS, Pisansky TM, Petersen IA, et al. Establishment of practice standards in nomenclature and prescription to enable construction of software and databases for knowledge-based practice review. *Pract Radiat Oncol.* 2016;6:e117-e126.
7. Mayo CS, Deasy JO, Chera BS, Freymann J, Kirby JS, Hardenberg PH. How Can we effect culture change toward data-driven medicine? *Int J Radiat Oncol.* 2016;95:916-921.
8. Yu J, Straube W, Mayo C, et al. Radiation therapy digital data submission process for national clinical trials network. *Int J Radiat Oncol Biol Phys.* 2014;90:466-467.
9. Santanam L, Hurkmans C, Mutic S, et al. Standardizing naming conventions in radiation oncology. *Int J Radiat Oncol Biol Phys.* 2012;83:1344-1349.
10. Denton TR, Shields LBE, Hahl M, Maudlin C, Bassett M, Spalding AC. Guidelines for treatment naming in radiation oncology. *J Appl Clin Med Phys.* 2016;17:123-138.
11. Mayo CS, Moran JM, Bosch W, et al. AAPM TG-263: Standardizing nomenclatures in radiation oncology. *Int J Radiat Oncol Biol Phys.* 2017;100:1057-1066.
12. Hamdy FC, Donovan JL, Lane JA, et al. 10-Year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. *N Engl J Med.* 2016;375:1415-1424.
13. Parker C, Clarke N, Logue J, et al. RADICALS (Radiotherapy and Androgen Deprivation in Combination after Local Surgery). *Clin Oncol.* 2007;19:167-171.
14. Roach MI. RTOG Clinical Trials Study Number 0924. Available at: https://www.rtog.org/ClinicalTrials/ProtocolTable/StudyDetails.aspx?study=0924. Accessed November 27, 2017.
15. Christiaens M, Collette S, Overgaard J, et al. Quality assurance of radiotherapy in the ongoing EORTC 1219-DAHANCA-29 trial for HPV/p16 negative squamous cell carcinoma of the head and neck: Results of the benchmark case procedure. *Radiother Oncol.* 2017;123:424-430.
16. R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2017.
17. Thompson RF. *RadOnc: Analytical Tools for Radiation Oncology*; 2017. Available at: https://CRAN.R-project.org/package=RadOnc. Accessed February 12, 2018.
18. Forde E, Kneebone A, Bromley R, Guo L, Hunt P, Eade T. Volumetric-modulated arc therapy in postprostatectomy radiation therapy patients: A planning comparison study. *Med Dosim.* 2013;38:262-267.
19. Caine H, Whalley D, Kneebone A, McCloud P, Eade T. Using individual patient anatomy to predict protocol compliance for prostate intensity modulated radiation therapy. *Med Dosim.* 2016;41:70-74.
20. Whalley D, Hruby G, Alfieri F, Kneebone A, Eade T. SpaceOAR hydrogel in dose-escalated prostate cancer radiotherapy: Rectal dosimetry and late toxicity. *Clin Oncol (R Coll Radiol).* 2016;28:e148-154.
21. D'Souza J, Ng V. Sieve-based entity linking for the biomedical domain. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing.* Vol 2. Stroudsburg, PA: Association for Computational Linguistics; 2015:297-302. Short Papers.
22. Bodenreider O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* 2004;32(Database Issue):D267-D270.
23. Soğancõoğlu G, Öztürk H, Özgür A. BIOSSES: A semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics.* 2017;33:i49-i58.
24. Mayo CS, Kessler ML, Eisbruch A, et al. The big data effort in radiation oncology: Data mining or data farming? *Adv Radiat Oncol.* 2016;1:260-271.
25. Ruan D, Shao W, Wong J, et al. SU-F-T-102: Automatic curation for a scalable registry using machine learning. *Med Phys.* 2016;43:3485.
26. Gallego B, Walter SR, Day RO, et al. Bringing cohort studies to the bedside: framework for a "green button: to support clinical decision-making. *J Comp Eff Res.* 2015;4:191-197.