Data collection of patient outcomes: one institution’s experience

Thomas J. Whitaker1,*, Charles S. Mayo2, Daniel J. Ma1, Michael G. Haddock1, Robert C. Miller3, Kimberly S. Corbin1, Michelle Neben-Wittich1, James L. Leenstra1, Nadia N. Laack1, Mirek Fatyga4, Steven E. Schild4, Carlos E. Vargas4, Katherine S. Tzou3, Austin R. Hadley3, Steven J. Buskirk3 and Robert L. Foote1

1Department of Radiation Oncology, Mayo Clinic, Rochester, Minnesota, USA
2Department of Radiation Oncology, University of Michigan, Ann Arbor, Michigan, USA
3Department of Radiation Oncology, Mayo Clinic, Jacksonville, Florida, USA
4Department of Radiation Oncology, Mayo Clinic, Phoenix, Arizona, USA

*Corresponding author. Department of Radiation Oncology, Mayo Clinic, 200 First St. S.W., Rochester, MN, USA. Tel: +01-507-255-2129; Fax: +01-507-284-0079; Email: Whitaker.Thomas@mayo.edu
(Received 30 November 2017; revised 2 February 2018; editorial decision 7 February 2018)

ABSTRACT

Patient- and provider-reported outcomes are recognized as important in evaluating quality of care, guiding health care policy, comparative effectiveness research, and decision-making in radiation oncology. Combining patient and provider outcome data with a detailed description of disease and therapy is the basis for these analyses. We report on the combination of technical solutions and clinical process changes at our institution that were used in the collection and dissemination of this data. This initiative has resulted in the collection of treatment data for 23 541 patients, 20 465 patients with provider-based adverse event records, and patient-reported outcome surveys submitted by 5622 patients. All of the data is made accessible using a self-service web-based tool.

Keywords: patient outcomes; big data; informatics; analytics; randomized controlled trials

INTRODUCTION

The clinical workflow in Radiation Oncology treatment generates ample opportunities for collecting patient outcome data. Once collected, there are many uses for patient outcome data. Focusing on scientifically rigorous quality and safety analyses would improve outcomes and cost-effectiveness [1]. Consistent clinical collection of patient-reported outcome (PRO) data facilitates early interventions that are targeted to the specific needs and concerns of patients [2], which in turn may result in improved outcomes and future treatment [3]. Furthermore, analysis of routine clinical data may facilitate cost-effectiveness analyses to objectively determine the value of new technologies or new therapy techniques [4, 5]. Pairing clinical data with machine learning offers the potential for finding new insights by combining a large variety of data types that would be too confounding in the context of other analysis techniques [6]. It is also possible to use clinical data to adopt a model-based approach when validating treatment efficacy, rather than rely on randomized controlled trials. However, successful implementation of the above vision relies on rapid-learning techniques, which require collecting large amounts of data, learning from the data, making modifications to therapeutic techniques and then iterating the process [7]. The success of this vision relies on high-quality data.

High-quality datasets that will meet the needs of the future have the following properties: they are composed of a variety of data, their data are structured, and they contain patient outcome measurements. To create high-quality datasets, a combination of technical solutions and clinical process changes are needed. These concepts are thoughtfully articulated by Mayo et al. [8, 9]. The datasets for the future contain detailed descriptions of a patient’s disease that are well connected to the treatment they received. The
treatment data needs to be entered in a structured manner so that it can be aggregated properly \cite{10-12}. The patient must then be clinically followed and their outcomes must be consistently documented. These are examples of the culture change required. It cannot be assumed that this data is completely documented in the medical record. A desire to faithfully and completely enter data that can be aggregated is something that must be cultivated.

Once the data is in the medical record, technology can aid by extracting and formatting the data for routine analysis, and making it accessible to users. Examples of technical solutions for data extraction and formatting can be found in Roelofs \textit{et al.} \cite{13} and similar works \cite{14, 15}. Technical solutions can also aid in collecting data that is not collected in a medical record system. This work outlines the clinical and technical solutions our institution went through in building our patient outcome dataset and the result of our efforts.

\section*{METHODS}

The design goals that guide our collection of patient outcomes data are: data should be collected systematically for all curative patients, both physician-reported \cite{16} and patient-reported adverse events (AEs) should be collected at standard time-points, and data should be available to users through a self-service interface. To achieve these goals, we have combined technical solutions with changes in clinical processes.

\subsection*{Technical solutions}

\textbf{Data aggregation}

Our institution’s radiation oncology services are provided at three main campuses, with two of the main campuses being composed of multiple regional facilities. A data warehouse was built to aggregate patient-outcome data for all patients across all facilities (Fig. 1). The data warehouse provides a single repository for data from multiple source systems. The hospital registration system is the source of detailed demographic data such as: age, race, ethnicity, date of birth, date of death, and residency location. The radiation oncology record-and-verify system is the source for: diagnosis, staging, pathology, provider-reported AEs (AEs), treatment dates, and provider data. An internally developed survey collection system is our source for all PRO data. Each main campus has its own instance of the record-and-verify system, and its version might differ from that of the other main campuses. Only two of the three main campuses share the same electronic-medical-record vendor. It is also important to anticipate that source systems for data will change over time. The warehouse makes delivery of data to the end user independent of these variations.

\textbf{Data accessibility}

Designing a tool for securely accessing data requires knowing who the users are and how they will use the data. Three major user groups were identified at our institution. The first group is the clinical management team. The clinical management team is interested in data that describe the entire patient population served by the department. For example: how many patients are treated per quarter, what were their diagnoses, and what modality did they receive? Clinical managers are typically interested in all patients treated in the clinic within a certain time period or with a certain modality rather than a particular diagnosis. The second user group is retrospective researchers. This user is interested in a special cohort. For example, patients treated for nasal cavity and paranasal sinus cancers

Fig. 1. The data input streams and the data output methods used when constructing our data warehouse. The record and verify box represents three separate instances of varying versions across three main campuses. The survey collection system is an internally built web-based tool for collection of patient-reported outcome surveys. Self-service data extraction is available via a web-based data-query tool. Programatic data extraction is also available for use by trial databases.
using intensity-modulated radiotherapy versus intensity-modulated proton therapy. This user knows what characteristics define the cohort, but they don’t know the specific patients in the cohort before data extraction takes place. Users working on model-based approaches are in this group. They use routine clinical data from well-defined cohorts to draw scientific insights. The third group is the prospective-trial investigators. They know before data extraction takes place which patients are in their cohort of interest. The trial eligibility criteria and treatment protocol precisely define the cohort.

Each user group must be able to query and extract data in a way that meets their needs. The technical solution at our institution was to provide two interfaces to meet the needs of all three user groups. The first interface is a web-based query tool. The tool allows the user to select patients based on: the campus at which the patients received treatment, the treating physician, the dates of treatment, and the primary and secondary diagnoses. After setting the selection criteria, the user can then view the resulting data within the tool. There are options that allow the user to save their selection criteria and recall them for future use or editing. The user is also allowed to extract the results in spreadsheet format. This interface has worked well for users of the first and second types described above. For users who know the specific patients in their cohort, the web tool allows them to use their patient list as a filter. This gives these users the ability to extract data for just their patients of interest. The second interface is a web service that allows an end user to programmatically call for particular data for a specific patient and receive the return as an xcm string. This interface allows a trial database to programmatically download data for a specific patient. The user benefits by reducing the quantity of data manually entered into trial databases.

Security of the data is a concern regardless of the user group. Our data extraction tool is integrated into our institution’s user authentication infrastructure. This ensures that only users with access to the source systems have access to the aggregated data.

**Patient-reported outcomes collection**

A PRO collection and management system, called PROSS, was developed to aid in the collection of patient-reported outcomes data. Every adult curative-intent patient is asked to take the PROMIS-10 quality of life survey. In addition, patients are asked to take an additional one to three surveys, depending on the patient’s site of disease. Within PROSS, the PROMIS-10 and the disease site-specific surveys are grouped to form a survey series. The survey series are given specific names, based on the disease site, to aid in the ordering process. The same concept is used for our pediatric population, but rather than being grouped by disease site, the survey content is grouped by age. PROSS allows a user to schedule surveys for a patient by selecting the appropriate survey series and entering the patient’s first and last days of treatment. The treatment dates are used to schedule delivery of surveys to the patient over the course of 5 years. The delivery time-points are: pre-treatment (pre-tx), post-treatment (post-tx), 3 months post-treatment (3 mon. post-tx), 6 months post-treatment (6 mon. post-tx), and annually. Once scheduled, the patient is immediately able to take the survey using an electronic tablet. If the patient does not take the surveys while in the department, an email is sent to the patient allowing them to take the survey via a web portal. If the patient does not respond to the email, they receive a phone call from a nurse, nurse practitioner, or physician assistant as a reminder and/or to provide help answering the survey over the phone. If the patient does not take the survey within a pre-defined time interval, the surveys are canceled. Each time-point has its own pre-defined time interval, typically half-way to the next time-point. Two weeks before the start of each time-point is due, PROSS activates the survey content. Activation makes the surveys available to be taken via electronic tablet and the web portal. Other features in PROSS allow survey delivery to be suspended if the patient decides to stop participating and also checks for a record of the patient’s death before sending email notices. Survey series activation, cancellation, and patient status monitoring are all automated processes that do not require further user interactions.

**Clinical processes**

Changes and additions to clinical processes are often left out when discussing outcomes data collection initiatives. However, clinical processes are the key to creating a culture of data collection [8]. It is the change in culture that results in rich and valuable data that ensures the success of each user’s endeavor. Our institution implemented changes to clinical process in three areas: standardization of nomenclature, recording of complete and detailed description of disease (diagnosis, pathology and staging), and systematic recording of AEs.

**Standardization of nomenclature**

A large portion of the data within an electronic medical record or a record-and-verify system is free text. The time needed to manage variation in free text can be a limiting factor on the size of an assembled dataset. By defining a standard nomenclature, this burden can be reduced substantially.

Defining standard target and critical structure definitions aids in being able to quickly aggregate dose–volume histogram data. Our institution currently uses nomenclature described by Mayo et al. [10]. Other nomenclatures have also been established, such as that of Santanum et al. [11]. The report from the American Association of Physicists in Medicine (AAPM) Task Group 263—Standardizing Nomenclature for Radiation Oncology has recently been approved and is in the process of being published. The nomenclature was collaboratively developed with members of the American Society of Therapeutic Radiation Oncology (ASTRO) and the European Society of Therapeutic Radiation Oncology (ESTRO). The report recommendations are anticipated to be widely adopted.

Description of a patient’s disease is another area in which the clinic is challenged by free text. The parsing of clinical notes that commonly hold the description of a patient’s disease, has led to rapid growth in the field of natural language processing. One possible solution is to change the clinical process so that this data is entered in a structured format. Our providers carefully document the patient’s primary and secondary diagnosis in the record-and-verify system. They also document the patient’s stage, pathology, and any other important descriptive information (i.e. HER2 status, Gleason scores, HPV status, etc.). After entering this data, the provider connects it to the course of treatment to which it applies.
Compliance in data entry is facilitated by weekly peer-to-peer new patient chart reviews. During this review, the diagnosis, pathology and staging entry are reviewed and corrected for errors and missing data. This consistent and repetitive peer review has been successful in producing a culture of data entry and curation.

Provider-reported adverse events

Selections from the CTCAE version 4.03 [16] AE lists have been created for each disease site. Using features within our record-and-verify system, we have created an easy-to-select list for each disease site subset. Providers use the list to document the presence or absence of the AEs in the list. Providers also document any AE the patient experiences that is not listed within the disease-specific subset. Providers document attribution for any AE of Grade 1 or above. The AEs are systematically recorded for all patients’ pre-treatment, post-treatment, at all follow-up visits, and when any change in status is noted.

Patient-reported outcomes

A systematic clinical process was put in place to maximize our patient-reported outcomes collection rate. The first step in the process is for the provider to clearly order a survey series for the patient. This order contains the specific survey series name the patient is to receive. Using the specific name of the survey series reduces confusion and errors when a radiation therapist schedules the surveys at the time of CT simulation. After CT simulation, patients have a scheduled education session with a research coordinator. At this session, the research coordinator provides education concerning the surveys and collects the pre-tx surveys using an electronic tablet. Note that the collection of surveys is part of the patient’s routine clinical care. The decision to use a research coordinator for education and pre-tx collection is based on their training and expertise in providing patient education. Two back-up methods are used in the event that the pre-treatment survey was not collected during the education session. The first back-up method is an email sent to the patient on their first day of treatment. The second back-up method occurs when the desk staff place the patient in an exam room for the patient’s first treatment visit with a provider. In the event that the patient has not completed their surveys, the desk staff will provide the patient with an electronic tablet to do so. The first attempt for collecting the post-treatment surveys is performed by the desk staff when placing the patient in a room for their last treatment visit. Again this is done using an electronic tablet. An email sent to the patient is used as the back-up method for collecting the post-treatment surveys. For follow-up surveys, an email to the patient is the primary collection method. A phone call from a nurse is the back-up method for follow-up time-points.

The variety of clinical roles used to collect PRO data is intentional. Giving every team member a role in the data collection process reinforces the importance of the data and builds up the culture of data collecting. Figure 2 illustrates the PRO collection workflow described above.

Fig. 2. The PRO data collection workflow. The workflow illustrates the intertwining of technical solutions and the clinical processes that make data collection initiative successful.
RESULTS AND DISCUSSION

Our institution started implementing the technical and clinical process described above on 29 April 2013. The first phase of implementation focused on recording detailed descriptions of each patient’s disease. The second focus was regular and proactive collection of AEs data. In April 2015, collection of PRO data was implemented. Figure 3 summarizes the collection rate for each type of data.

The treatment data shown in Fig. 3 are for the patients who started treatment in the given quarter, and who were properly connected to detailed clinical data. This curve grew rapidly at first and then quickly plateaued. The shape is a direct reflection of the number of patients treated in a quarter. The small growth seen in later quarters is due to expanded services. In total, the data warehouse now holds treatment and disease information for 23,451 patients. These numbers are not exact treatment numbers. In order for the data to be automatically extracted from the record-and-verify system to the data warehouse, the treatment and disease description data must be properly connected.

The rate of recorded provider-reported AEs is shown as the darkest shade of gray in Fig. 3. The AE data reported in Fig. 3 is the number of patients who had an AE recorded in the particular quarter. This includes the recording of Grade 0 events (i.e. the absence of an AE). Figure 3 shows it took a little longer to achieve data entry compliance. However, it can also be seen that compliance has been achieved and that the rate of AE recording surpasses patients being treated. This indicates that AEs are also being recorded for patients receiving post-treatment follow-up care. Currently, 20,465 unique patients had AE data recorded in the data warehouse.

The last major implementation was for patient-reported outcomes. The PRO data is represented as the medium-gray-shaded distribution seen on the right-hand side of Fig. 3. The value reported in Fig. 3 is the number of patients that completed a PRO survey in that particular quarter. PRO implementation started with the prostate disease site at a single campus in April 2015. Shortly thereafter, all disease sites (at a single campus) were added. This stage of the implementation can be seen in the steep growth in survey data collected between the second quarter of 2015 and the first quarter of 2016. Once fully implemented at one campus, the implementation continued with the remaining campuses. The slow but steady growth seen following the first quarter of 2016 is illustrative of the difficulty in obtaining compliance in a large multicampus system. PRO data has been harder to collect than the other data types because it is a totally new data type that was not as extensively

![Fig. 3. Patients per quarter who had treatment data, provider-reported adverse event data, or patient-reported outcomes data recorded in a given quarter. The darkest-shaded region, the deepest layer, depicts the provider-reported adverse event collection numbers. The lightest-shaded region and the top layer is the collection rate of the treatment data. Sandwiched in between, in a medium shade of gray, is the patient-reported outcome collection rate.](image-url)
Fig. 4. The relative collection rate of patient-reported outcome data as a function of time relative to treatment. The completion rate is the percentage of the patients who took at least one of their scheduled surveys. The canceled rate is the percentage of the patients who did not take at least one of their scheduled surveys within the allowed time-frame.

collected previously. The expanded collection was a new source of clinical work rather than a change in clinical process. However, the steady growth in collection rate also indicates that diligent efforts are in the end rewarded. To date, PRO data has been collected for 5622 patients. Figure 4 shows our PRO collection rate for the various time-points. The completed column is the percentage of the patients who have completed at least one survey for the given time-point. Each time-point has a particular time-frame within which the patient must complete the surveys, at the end of which the surveys are canceled. The canceled column represents the number of patients who failed to complete at least one survey for the given time-point. Since patients are asked to take more than one survey, it is possible, though infrequent, that a patient appears in both the completed and canceled columns. The percentage is only for patients who have surveys activated for the particular time-point. Percentages less than 100% when the completed and canceled columns are summed represent the percentage of surveys that are active for collection, but have not been collected.

CONCLUSION
For a models-based approach to be used for all disease sites in radiation oncology, systematic collection of all clinical data is needed. Our institution has shown that this goal can be achieved. Clinical data, provider AE data, and patient-reported outcome data can be collected for all disease sites if a strategy of using both technical solutions and clinical-process changes is adopted.

REFERENCES
1. Potters L, Ford E, Evans S et al. A systems approach using big data to improve safety and quality in radiation oncology. Int J Radiat Oncol Biol Phys 2016;95:885–9.
2. Sloan JA, Halyard M, El Naqa I et al. Lessons from large-scale collection of patient-reported outcomes: implications for big data aggregation and analytics. Int J Radiat Oncol Biol Phys 2016;95:922–9.
3. Basch E, Deal AM, Dueck AC et al. Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. JAMA 2017;318:197–8.
4. Richard P, Phillips M, Smith W et al. Cost-effectiveness analysis of intensity modulated radiation therapy versus 3-dimensional conformal radiation therapy for preoperative treatment of extremity soft tissue sarcomas. Int J Radiat Oncol Biol Phys 2016;95:999–1008.
5. Zemplenyi AT, Kalo Z, Kovacs G et al. Cost-effectiveness analysis of intensity-modulated radiation therapy with normal and hypofractionated schemes for the treatment of localised prostate cancer. Eur J Cancer Care 2018;27:e12430.
6. Bibault JE, Giraud P, Burgun A. Big Data and machine learning in radiation oncology: state of the art and future prospects. Cancer Lett 2016;382:110–7.
7. Lambin P, Roelofs E, Reymen B et al. ‘Rapid Learning health care in oncology’—an approach towards decision support systems enabling customised radiotherapy. Radiother Oncol 2013;109:159–64.
8. Mayo CS, Deasy JO, Chera BS et al. How can we effect culture change toward data-driven medicine? Int J Radiat Oncol Biol Phys 2016;95:916–21.
9. Mayo CS, Kessler ML, Eibbruch A et al. The big data effort in radiation oncology: data mining or data farming? Adv Radiat Oncol 2016;1:260–71.
10. 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:e117–26.
11. Santanam L, Hurkmans C, Mutic S et al. Standardizing naming conventions in radiation oncology. Int J Radiat Oncol Biol Phys 2012;83:1344–9.
12. Sioutos N, de Coronado S, Haber MW et al. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform 2007;40:30–3.
13. Roelofs E, Dekker A, Meldolesi E et al. International data-sharing for radiotherapy research: an open-source based infrastructure for multicentric clinical data mining. Radiother Oncol 2014;110:370–4.
14. Jochems A, Deist TM, van Soest J et al. Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital—a real life proof of concept. Radiother Oncol 2016;121:459–67.
15. Zhang L, Hub M, Mang S et al. Software for quantitative analysis of radiotherapy: overview, requirement analysis and design solutions. Comput Methods Programs Biomed 2013;110:528–37.
16. National Cancer Institute (U.S.). Common Terminology Criteria for Adverse Events v4.03 (CTCAE v4.03). Bethesda, MD: U.S. Dept. of Health and Human Services, National Institutes of Health, National Cancer Institute, 2009. https://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm.