19-1661

Patient-level validation of prostate cancer data collected via automated extraction from structured and unstructured electronic health record (EHR) records

Matthew R. Cooperberg, William Meeks, Ji Qi, Rodney Dun, Sanyog Pendharkar, Daniel Pichardo, Anna Johnson, Susan Linsell, Raymond Fang, Steven Schlossberg, James E. Montie

INTRODUCTION AND OBJECTIVES: The AUA Quality (AQUA) Registry now includes data on _ patients managed by _ urologists across the country. AQUAs databases are populated by automated extraction of data from a variety of EHR systems. Some data (e.g., billing codes and orders) usually exist as structured data in EHRs. Others (e.g., cancer grade) usually do not, and must be identified via regular expression or the use of natural language processing. As a test of data extraction quality, we performed a patient-level validation of prostate cancer data from two AQUA practices compared to the manually abstracted data available through their participation in the Michigan Urological Surgical Improvement Collaborative (MUSIC).

METHODS: Data were collected from men newly diagnosed with prostate cancer between 2014 and 2017 at two urology practices in Michigan. AQUA data were extracted using EHR connector software (FIGMD Inc, San Diego, CA), and MUSIC data were manually abstracted by trained staff at each site with annual onsite quality audits. Date of diagnosis, Gleason score (primary and secondary), diagnostic PSA, number of biopsy cores (positive and total), clinical staging, and primary treatment were compared. Percent of cases with missing information on each variable was also evaluated for both registries.

RESULTS: A total of 725 patients from the two practices were linked between AQUA and MUSIC registry. The rate of missing data in each registry as well as matching rates for values when identified are shown in Table 1. The most common mismatches for treatment were between brachytherapy and external-beam radiation, and between radiation and primary androgen deprivation.

CONCLUSIONS: Automated extraction of both structured and unstructured data from EHRs is possible, and has the potential to substantially reduce the time and cost of disease registry population. Adjustments to algorithms will continually improve the quality of the automated abstraction.

Source of Funding: None

Table 1. AQUA MUSIC Data Validation Result

Variable	% Missing in AQUA	% Missing in MUSIC	% Match[1] when not missing	% Mismatch when not missing
Date of diagnosis	0.0	2.6	90.2	9.8
Gleason score, primary	9.9	2.7	94.1	5.9
Gleason score, secondary	9.9	2.7	84.0	16.0
Number of positive cores	20.1	3.0	82.9	17.1
Number of total cores taken	20.1	3.0	30.7	69.3
PSA value	28.0	3.1	53.4	46.6
Primary treatment type	21.5	23.3	79.0	21.0
Clinical T stage value	39.7	3.1	73.2	26.8

1. Definition of match is as follows: date of diagnosis is within 30 days apart; Gleason individual score is the same or Gleason sum is the same; number of biopsy cores is within 3 cores different; PSA value is within 20% different; Same treatment type; Same T stage number with same or different letter.