original report

Automating the Determination of Prostate Cancer Risk Strata From Electronic Medical Records

Justin R. Gregg
Maximilian Lang
Lucy L. Wang
Matthew J. Resnick
Sandeep K. Jain
Jeremy L. Warner
Daniel A. Barocas

Justin R. Gregg, **Maximilian Lang**, **Lucy L. Wang**, **Matthew J. Resnick**, **Jeremy L. Warner**, and **Daniel A. Barocas**, Vanderbilt University Medical Center; **Matthew J. Resnick** and **Jeremy L. Warner**, Vanderbilt University; **Matthew J. Resnick**, Tennessee Valley Veterans Administration Health Care System; and **Sandeep K. Jain**, Vanderbilt University School of Medicine, Nashville, TN.

J.L.W. and D.A.B. contributed equally to this work.

Supported in part (data collection and analysis) by National Center for Advancing Translational Sciences Grant No. UL1 TR000445, National Cancer Institute (NCI) Grant No. U24 CA194215, and the Vanderbilt-Ingram Cancer Center Support Grant/NCI Grant No. P30 CA068485.

Corresponding author: Justin R. Gregg, MD, A1302 Medical Center North, Vanderbilt University Medical Center, Nashville, TN 37232; e-mail: justin.r.gregg@gmail.com.

abstract

Purpose **Risk stratification underlies system-wide efforts to promote the delivery of appropriate prostate cancer care. Although the elements of risk stratum are available in the electronic medical record, manual data collection is resource intensive. Therefore, we investigated the feasibility and accuracy of an automated data extraction method using natural language processing (NLP) to determine prostate cancer risk stratum.**

Methods **Manually collected clinical stage, biopsy Gleason score, and preoperative prostate-specific antigen (PSA) values from our prospective prostatectomy database were used to categorize patients as low, intermediate, or high risk by D'Amico risk classification. NLP algorithms were developed to automate the extraction of the same data points from the electronic medical record, and risk strata were recalculated. The ability of NLP to identify elements sufficient to calculate risk (recall) was calculated, and the accuracy of NLP was compared with that of manually collected data using the weighted Cohen's κ statistic.**

Results **Of the 2,352 patients with available data who underwent prostatectomy from 2010 to 2014, NLP identified sufficient elements to calculate risk for 1,833 (recall, 78%). NLP had a 91% raw agreement with manual risk stratification ($\kappa = 0.92$; 95% CI, 0.90 to 0.93). The κ statistics for PSA, Gleason score, and clinical stage extraction by NLP were 0.86, 0.91, and 0.89, respectively; 91.9% of extracted PSA values were within $\pm$ 1.0 ng/mL of the manually collected PSA levels.**

Conclusion **NLP can achieve more than 90% accuracy on D'Amico risk stratification of localized prostate cancer, with adequate recall. This figure is comparable to other NLP tasks and illustrates the known tradeoff between recall and accuracy. Automating the collection of risk characteristics could be used to power real-time decision support tools and scale up quality measurement in cancer care.**

INTRODUCTION

Virtually all clinical practice guidelines and quality measures for processes of care in oncology are cancer stage specific or risk stratum specific. Therefore, to determine whether evidence-based care is delivered to the appropriate candidate at the correct point in the course of his or her disease, one must know the cancer stage and other factors that comprise cancer risk. Although the delivery of health care services may be gleaned readily from claims data, cancer stage and risk are usually determined by examination of the medical record, a process that is often labor intensive and error prone. These key pieces of information are the basis for communication between researchers and clinicians; however, they remain buried deep within the electronic medical record (EMR), where they may be nevertheless accessible to automated extraction.

One example of the utility of quality measurement in oncology is the use of advanced imaging technologies among men with clinically localized prostate cancer. Most data demonstrate that such imaging is unnecessary in low-risk disease but indicated in high-risk cancers.[1,2] Accordingly, organizations such as the Physician Quality Reporting System and National Comprehensive Cancer Network publish risk stratum–specific recommendations for imaging.[3] Additionally, the American Society of Clinical Oncology recommends against the use of staging imaging in patients with low-risk prostate cancer as part of its Choosing Wisely campaign.[4] However, imaging use remains high among low-risk candidates,

leading to excessive resource use.[5] Efforts to reduce this use in low-risk patients have been remarkably successful but have required concerted data collection across practice sites, followed by data analysis and comparative performance feedback and/or decision support interventions.[6]

To scale up this approach, automated methods to characterize disease risk are attractive potential solutions. D'Amico risk classification is a validated risk stratification paradigm comprising clinical T stage (rectal examination result), immediate prediagnostic serum prostate-specific antigen (PSA) level, and prostate biopsy grade (Gleason score).[7] Thomas et al[8] recently showed that natural language processing (NLP) could be used to extract some of the clinical data needed for D'Amico risk classification, with 97.6% accuracy; however, only data found within pathology reports were extracted. Whether NLP can automatically and accurately place patients with prostate cancer into D'Amico clinical risk groups using information from across the EMR remains largely unknown.

To this end, we aimed to automate the collection of the components of prostate cancer risk stratification and verify the fidelity of that automation against that of manually abstracted data. We hypothesized that prostate cancer risk group could be accurately determined using extracted data in at least 90% of patients using NLP algorithms.

METHODS

With institutional review board approval, we gathered clinical risk stratum information (preoperative PSA, Gleason score, and clinical T stage) from all patients who underwent radical prostatectomy at Vanderbilt University Medical Center (VUMC) from 2010 through 2014. These data had previously been collected as part of a prospective institutional prostatectomy database, which functions as a local registry. Two trained abstractors with more than 15 years of clinical urologic experience were used for the creation of the prospective database.

NLP algorithms were developed to automate extraction of the D'Amico risk group elements. Content and metadata for clinical documents are stored in the VUMC research data warehouse and include progress notes, clinical communications, operative notes, pathology reports, and other documents generated within the VUMC system; outside scanned documents are not included.[9] Regular expressions were applied to the content strings of all clinical documents in a candidate patient's EMR dated from up to 1 year before prostatectomy until the day before prostatectomy for PSA and Gleason score (Appendix). T stage was extracted from cancer staging forms available within the EMR for optional use. Table 1 summarizes pseudocode describing the process of extraction for each of the three data elements. An iterative process was used whereby the algorithm output was evaluated against 10 randomly selected records by a subject matter expert (J.L.W.). This process was then continued until the algorithm achieved more than 90% accuracy.

We conducted a preliminary analysis of the prevalence of the key phrases "low risk," "intermediate risk," and "high risk" in clinical notes up to 1 year before prostatectomy for included patients. A minority of the records (22%) had any of these key phrases present, and some (9%) had two or more of the key phrases during the examined time period. D'Amico risk group[7] was therefore assigned on the basis of existing manually extracted data (low [PSA $\leq$ 10 ng/mL, Gleason score $\leq$ 6, and clinical stage $\leq$ T2a], intermediate [PSA $>$ 10 to $\leq$ 20 ng/mL, Gleason score $\leq$ 7, and clinical stage $\leq$ T2b], or high risk [PSA $>$ 20 ng/mL, Gleason score $>$ 8, and clinical stage $>$ cT2c]). Risk group was then assigned on the basis of data extracted using NLP and compared with the manual extraction–based risk group. Risk group was still assigned in the case of missing data elementsa as long as the elements present were sufficient (eg, having PSA $>$ 20 ng/mL or Gleason score $>$ 8 or T stage $>$ cT2c was sufficient to assign high risk). Raw agreement (percentage of successful NLP patient extractions that resulted in correct D'Amico risk group classification) was calculated. The weighted κ statistic[10] and 95% CI were then used to measure agreement between manually extracted and NLP-determined risk group for the individual characteristics. In an exploratory analysis, we determined what proportion of patients with incalculable risk group by NLP had sufficient data available to determine whether they were at least intermediate risk.

We hypothesized that the proportion of patients who were accurately risk stratified using NLP was at least 0.9. The estimated sample size needed for comparison of this proportion with the alternative hypothesis that the proportion was actually 0.85 was 362 patients, with a probability (power) of 0.9. The type I error probability associated with this test was 0.05 (one sided).

Finally, we conducted a manual discrepancy analysis of a convenience sample of the disagreements. A clinical subject matter expert (S.K.J.) reviewed

Table 1. Pseudocode Describing NLP Algorithm

Steps
A. Determine preoperative PSA
Step 1. Extract PSA from laboratory-based scores at VUMC
Retrieve all results prior to and within 1 year of date of surgery
Capture the highest result as the preoperative PSA
Step 2. Extract PSA results from clinical notes
Retrieve strings following PSA, including "undetectable," or "<,0.10," or numeric strings not in date format, ignoring punctuation, white space, and key words such as "value, score, rose, rising, remains, of, to, was, is, =, at" between "PSA" and result
Date stamp each extracted result with the date that the note was recorded in the EMR
When a result is found more than once, capture earliest date stamp as test date for result
Retain all results before and within 1 year of date of surgery
Capture the highest result as the preoperative PSA
Step 3. Report preoperative PSA
If result found in step 1, use it
If not, use result found in step 2
If no result found, report result as unknown
B. Calculate biopsy-based Gleason scores
Step 1. Extract Gleason scores from VUMC pathology reports written before date of surgery
Retrieve all scores written as "primary + secondary" or "primary + secondary = total"
Calculate total score as "parsed primary score + parsed secondary score"
Parse out the highest total or calculated total score(s)
Step 2. Extract Gleason scores from other clinical notes written before date of surgery
Retrieve all scores written as "primary + secondary" or "primary + secondary = total" in various forms
Calculate total score as "parsed primary score + parsed secondary score"
Parse out the highest total or calculated total score(s)
Step 3. Extract total Gleason scores from all clinical notes written before date of surgery
Retrieve scores written as "total" only in various forms
Parse out the highest total score
Step 4. Report biopsy-based Gleason score
If score found in step 1, use it
If not found in step 1, use score found in step 2
If not found in step 2, use score found in step 3
If no score found, report score as unknown
C. Extract clinical stage from cancer staging forms
Search record for cancer staging (version 7; prostate) forms
Retrieve the latest clinical stage from forms
If not present, report clinical stage as unknown

Abbreviations: EMR, electronic medical record; NLP, natural language processing; PSA, prostate-specific antigen; VUMC, Vanderbilt University Medical Center.

the EMRs of selected patients for whom staging was changed between data sets to determine whether the error was from the manual abstraction or from the automated extraction. This determination was adjudicated by a second subject matter expert (J.L.W.), and disagreements were resolved through consensus discussion among all authors.

RESULTS

A total of 2,352 patients underwent prostatectomy during the time period, all of whom had at least some data available for analysis. Average age was 61.7 years (standard deviation, 7.1 years), and 90.0% were white. Average preoperative PSA value was 6.6 ng/mL (standard deviation, 4.7 ng/mL). Table 2 summarizes gold-standard D'Amico risk group classification calculated on the basis of the manually collected prospective database.

The NLP algorithms identified at least one of the required data elements in 2,351 patients (99.95%). NLP identified a combination of PSA, Gleason score, and clinical T stage sufficient to calculate D'Amico risk stratification in 1,833 patients (recall, 78.0%). Table 3 summarizes D'Amico risk group classification success of NLP compared with manual data collection. Raw agreement (precision) was 91.0%. Weighted κ for group classification was 0.92 (95% CI, 0.90 to 0.93). Of the 505 patients with incalculable D'Amico risk stratification by NLP, 219 (43.4%) had sufficient data available to calculate that they were at least intermediate risk, for a modified recall of 87.2%.

Table 4 summarizes NLP extraction of the individual PSA, Gleason score, and clinical T-stage components compared with those obtained by manual collection. Weighted κ statistics were 0.86 (95% CI, 0.82 to 0.90), 0.91 (95% CI, 0.90 to 0.93), and 0.89 (95% CI, 0.85 to 0.94) for PSA, Gleason score, and clinical T-stage categories, respectively. A total of 2,038 (91.9%) of 2,218 extracted PSA values were within $\pm$ 1.0 ng/mL of manually collected values.

The results of the manual discrepancy analysis of 10 patient cases are listed in Table 5. In six of nine patient cases, the manual review agreed with the NLP-determined risk stratum; in three patient cases, the manual review agreed with the original gold-standard risk stratum; and in one patient case, the risk stratum did not change despite the discrepancy.

DISCUSSION

We developed an algorithm using NLP for the automated extraction of clinical data required for risk stratification of patients with clinically localized prostate cancer. After NLP-driven risk

Table 2. D'Amico Risk Group Stratification of Study Cohort As Determined by Manual Data Extraction

Risk Group	No. of Patients (%)
Low	931 (39.6)
Intermediate	830 (35.3)
High	531 (22.6)
Could not be determined	60 (2.6)

stratification, the success of extraction and accuracy were compared with those of clinical data obtained from our prospective institutional prostatectomy database. We demonstrate a recall value of 78% for patients with available data, of whom 91% were accurately risk stratified. In our exploratory analysis, we found that recall increased to 87% when including a determination of at least intermediate risk for patients with incalculable exact risk strata. Discrepancies, when present, tended to be relatively minor; major discrepancies (stratifying as high risk when the patient should have been low risk) only affected nine (0.5%) of 1,833 patients. Notably, the recall value was underestimated, because the gold standard did not contain sufficient data to calculate D'Amico risk in 60 patient cases; nevertheless, we were able to classify 14 of these using the NLP algorithm (data not shown). The successful development of NLP tools to extract data surrounding extent of disease in oncology has broad implications for providers and for health systems transitioning to a value-based health care environment.

A manual review of a sample of discrepancies revealed that the NLP algorithm was more often correct than the gold standard in these cases. Although this finding might not extrapolate to all discrepancies, the review did reveal some important themes. Many of the discrepancies involved the not-uncommon situation where a local pathology review disagreed with the original pathology report (either in the direction of higher or lower maximum grade). Discrepancies in pathologist-assigned Gleason score are known to occur and can affect risk stratification.[11] For this analysis, we took the VUMC-assigned Gleason score as the correct score; it is possible that the original manual abstractors may have instead used a different rule, which led to the discrepancy. This finding serves to underscore the importance of institutional pathology review and should be accounted for in future use and expansion of the algorithm.

Traditionally, clinical risk strata are collected for use in individual patient counseling and decision making. Clinical risk grouping has a clear impact on prognosis[12] and affects disease treatment recommendations.[13] With a growing national focus on health care cost and quality, risk-based quality measures have been developed as indicators of the value of care.[14] Pay-for-performance systems, such as the Physician Quality Reporting System, use several quality measures relating to prostate cancer care, including the avoidance of bone scans in patients diagnosed with low-risk disease.[15] Although investigations in other disciplines suggest that quality indicator adherence may not be associated with improved outcomes,[16,17] they are considered benchmarks of high-quality care.[15] As we shift from a fee-for-service model to value-based care,[18] and Medicare payments become increasingly tied to quality or value via alternative payment models,[19] physician performance on such measures is of increasing importance to both clinicians and payers. In fact, health systems, hospitals, private practices, and specialty societies are investing in or hiring vendors to extract data from EMRs to demonstrate compliance and avoid financial penalties.[20] Efficient and accurate risk grouping is thus a prerequisite for comprehensive evaluation of and compliance with quality measures.

The determinants of prostate cancer risk stratification are often buried within text-based medical records. Examinations of large cohorts of patients and calculations of risk-related outcomes traditionally require manual record abstraction. We were fortunate to take advantage of an existing prospectively collected database, making a comparison between manual abstraction time and NLP development effort difficult; databases created expressly for the purposes of training NLP algorithms can have significant time costs. It is clear that with the massive number of narrative data now generated by EMRs, human abstractors must concentrate on high-value documents, often characterized by idiosyncratic hunting and gathering activities, whereas NLP can be applied to all documents. NLP simplifies and standardizes the data extraction process, potentially obviating the need to hire employees to manually extract risk data from patient records and allowing already

Table 3. D'Amico Risk Group As Determined by Manual Data Collection and NLP

Risk Group by Manual Data Collection	Risk Group by NLP		
	Low	Intermediate	High
Low	614	57	9
Intermediate	41	570	40
High	0	18	484

Abbreviation: NLP, natural language processing.

Table 4. Risk Stratification Success by Individual Strata

Manual Data Collection	NLP		
PSA	⩽ 10	> 10 to ⩽ 20	> 20
⩽ 10	1,886	23	15
> 10 to ⩽ 20	15	211	5
> 20	3	7	53
Gleason score	⩽ 6	7	8-10
⩽ 6	986	81	4
7	53	880	20
8-10	0	27	290
Clinical T stage	T1-T2a	T2b	T2c-T3
T1-T2a	1,625	8	11
T2b	8	49	0
T2c-T3	3	1	71

NOTE. Categories are those used to assign points in the D'Amico risk stratification algorithm.
Abbreviations: NLP, natural language processing; PSA, prostate-specific antigen.

employed abstractors to refocus their work on quality improvement and quality assurance activities (although manual abstraction would still be required for the approximately 20% of patient cases not calculable by the current algorithm). NLP also provides the benefit of automated extraction without interrupting the clinical workflow, because it is performed on the back end of clinical documentation. Alternative methods of data extraction include the use of electronic forms or pop-ups, which require active physician input before automated extraction. Although accurate, these forms require clinician time and additional interaction with the EMR, two of the top four contributors to physician dissatisfaction in a recent multispecialty survey.[21] Work by investigators at the University of Michigan shows that much of the clinical risk stratification information needed for patients with prostate cancer can automatically be obtained from pathology reports; however, these forms often do not include clinical examinations or laboratory values and are tailored to the format of individual health system reports. Synoptic structured pathology reports are not currently mandated for prostate biopsy specimens, and as such, most reports will be variable to some degree. Although the Health Information Technology for Economic and Clinical Health Act and related rules for Certified Electronic Health Record Technology lay the groundwork for an increasingly structured clinical record, the fact remains that many of the elements needed here will not be mandated as structured in the foreseeable future. For example, PSA is frequently not rechecked at

Table 5. Discrepancy Analysis

Patient Case	Risk			Notes
	NLP	Manual Review (gold standard)	Clinical Review	
1	Intermediate	Low	Intermediate	Discrepancy in Gleason score and PSA; manual review used OSH pathology Gleason (3 + 3 = 6), which was upgraded at VUMC to 3 + 4 = 7
2	Intermediate	Low	Intermediate	Discrepancy in Gleason score; manual review used OSH pathology Gleason (3 + 3 = 6), which was upgraded at VUMC to 3 + 4 = 7
3	High	Low	Low	Discrepancy in Gleason score; pathology note details why this patient case should be treated as 3 + 3 = 6 as opposed to 4 + 4 = 8; NLP algorithm parsed 4 + 4 = 8 from this discussion
4	Intermediate	Low	Intermediate	Discrepancy in PSA; clinical note: "prebiopsy PSA of 20 but then decreased to 5"
5	Low	Intermediate	Low	Discrepancy in Gleason score; manual review used OSH pathology Gleason score (4 + 3 = 7), which was downgraded at VUMC to 3 + 3 = 6
6	Low	Intermediate	Low	Discrepancy in Gleason score; manual review used OSH pathology Gleason score (3 + 4 = 7), which was downgraded at VUMC to 3 + 3 = 6
7	High	Intermediate	High	Discrepancy in Gleason score; manual review used OSH pathology Gleason (4 + 3 = 7), which was upgraded at VUMC to 4 + 4 = 8
8	High	High	High	Inconsequential discrepancy in Gleason score; VUMC pathology: 3 + 5 = 8; OSH pathology: 4 + 3 = 7
9	Intermediate	High	High	Discrepancy in Gleason score; NLP erroneously pulled 4 + 3 = 7 instead of 3 + 5 = 8
10	Intermediate	High	High	Discrepancy in clinical T stage; clinical note mentioned "some bilateral induration of the prostate and a little bit of nodularity … impression: stage T2c carcinoma of the prostate," but clinical staging form recorded T1c, which is what NLP extracted

Abbreviations: NLP, natural language processing; OSH, outside hospital; PSA, prostate-specific antigen; VUMC, Vanderbilt University Medical Center.

tertiary care facilities, such that the only records of PSA are to be found in machine-readable text. Furthermore, structured data elements, such as International Classification of Diseases billing codes, can still be subject to high error rates and are often not sufficient for phenotyping activities.[22]

Provided physicians and their associated health systems have simplified access to data related to risk-stratified health measures, a question remains as to whether they can use these data to improve the quality of care provided. Investigators at the University of Michigan examined an intervention during which a broad consortium of urology practices received an educational intervention about the use of bone scan and computed tomography for low-risk prostate cancer. After the so-called clinical champion–driven intervention, a significant decrease in bone scan and computed tomography scan use was reported in the pooled consortium.[6] NLP-powered support tools and data reporting therefore offer the potential to dramatically increase the efficiency and speed at which risk-stratified quality information may be used to affect clinical care.

Any shift from manual extraction to automated data collection does, however, inherently result in a decreased level of precision in describing patient cohorts. There is no established standard for the minimum acceptable levels of precision and recall for research and/or operational needs, although precision greater than 90% is generally considered high. Aiming for higher levels of precision introduces the possibility of overfitting to what may not be an error-free data set, as was revealed during the discrepancy analysis (Table 5). It is possible to improve the recall of NLP using liberalized search and extraction of terms, increasing the sensitivity of the extraction. However, this reduces accuracy by applying less stringent definitions to test extraction, thereby limiting the specificity of the test. NLP advances such as neural networks or other machine-learning approaches could be used; however, strategies such as these are much more complex and difficult to implement than the text-search methods outlined in our study. Increased supervision of NLP algorithms can also help improve accuracy, although this comes at the cost of intensified reliance on human interaction and optimization.[23] It is likely that the reported rates of recall and accuracy are acceptable for the purposes of quality measurement and assessment; however, additional studies will be needed to demonstrate these benchmarks if NLP is expanded to more clinical practices, especially if health systems require more accurate recall than the 78% demonstrated in this cohort.

Our study is limited in that the algorithm was created and validated using a single-institution EMR for patients who ultimately underwent prostatectomy. Other forms of treatment for early-stage prostate cancer (eg, external-beam radiation) were intentionally excluded because we did not have prospectively collected data to serve as a benchmark for NLP output. Our algorithm was tailored to the specific format of the clinical documentation found in the electronic record of a presurgical patient and was not necessarily applicable to broader health systems or patient groups. Specifically, VUMC used the homegrown StarPanel EMR system at the time of this analysis. StarPanel is an advanced EMR that is well integrated with a research data warehouse called the Research Derivative[9]; this pre-existing ecosystem facilitated the organization and accessibility of data used by our NLP algorithm, including the use of clinical staging forms and electronic laboratory results. We acknowledge that such assets are not available at many other institutions, which may limit generalizability. Going forward, the algorithm will require upfront costs associated with modification and validation as it is expanded to additional health systems and patients. Whether these costs favorably compare with those of human abstractors has not been well studied, but large programs such as National Cancer Institute–SEER are actively evaluating NLP technologies, in part because of potential cost savings. Through our experience developing algorithms that use NLP, it is often the extraction of clinical stage strata that is the most challenging to complete accurately.[24]

Future directions for our work include further risk stratification beyond using only Gleason total score, PSA, and clinical stage. Risk scores such as the National Comprehensive Cancer Network risk grouping, which is used by the Centers for Medicare and Medicaid Services Oncology Care Model program,[25] take primary and secondary Gleason scores into account for finer gradations of risk categorization. As proposed by D'Amico et al,[26] risk may also be further stratified into favorable and unfavorable intermediate risk on the basis of percentage of positive prostate cores, a more complex parameter. An interesting area of future research could be to alter our NLP algorithm to collect and calculate these percentages from pathology reports.

In conclusion, in this study, we demonstrate that NLP can achieve greater than 90% accuracy on D'Amico risk stratification of localized prostate

cancer. Recall was 78% using this technology at a single institution, and this performance is considered to be acceptable for clinical NLP tools. Validation and expansion of NLP algorithms may enable automated extraction of clinical risk characteristics to power decision support tools, disease registries, and quality measurement in cancer care.

DOI: https://doi.org/10.1200/CCI.16.00045
Published online on ascopubs.org/journal/cci on June 8, 2017.

AUTHOR CONTRIBUTIONS

Conception and design: Justin R. Gregg, Daniel A. Barocas
Administrative support: Matthew J. Resnick
Collection and assembly of data: Justin R. Gregg, Maximillian Lang, Lucy L. Wang, Sandeep K. Jain, Jeremy L. Warner
Data analysis and interpretation: Justin R. Gregg, Maximillian Lang, Matthew J. Resnick
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

Justin R. Gregg
No relationship to disclose

Maximilian Lang
No relationship to disclose

Lucy L. Wang
No relationship to disclose

Matthew J. Resnick
Consulting or Advisory Role: Janssen, MDxHealth
Research Funding: Genomic Health (Inst)

Sandeep K. Jain
No relationship to disclose

Jeremy L. Warner
No relationship to disclose

Daniel A. Barocas
Consulting or Advisory Role: AstraZeneca, Tolmar
Honoraria: AstraZeneca, Tolmar

REFERENCES

1. Carroll PR, Parsons JK, Andriole G, et al: Prostate cancer early detection, version 1.2014: Featured updates to the NCCN guidelines. J Natl Compr Cancer Netw 12:1211-1219, 2014
2. Eberhardt SC, Carter S, Casalino DD, et al: ACR appropriateness criteria prostate cancer: Pretreatment detection, staging, and surveillance. J Am Coll Radiol 10:83-92, 2013
3. Miller DC, Saigal CS: Quality of care indicators for prostate cancer: Progress toward consensus. Urol Oncol 27:427-434, 2009
4. American Society of Clinical Oncology: Ten Things Physicians and Patients Should Question. http://www.choosingwisely.org/societies/american-society-of-clinical-oncology/
5. Prasad SM, Gu X, Lipsitz SR, et al: Inappropriate utilization of radiographic imaging in men with newly diagnosed prostate cancer in the United States. Cancer 118:1260-1267, 2012
6. Ross I, Womble P, Ye J, et al: MUSIC: Patterns of care in the radiographic staging of men with newly diagnosed low risk prostate cancer. J Urol 193:1159-1162, 2015
7. D'Amico AV, Whittington R, Malkowicz SB, et al: Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. JAMA 280:969-974, 1998
8. Thomas AA, Zheng C, Jung H, et al: Extracting data from electronic medical records: Validation of a natural language processing program to assess prostate biopsy results. World J Urol 32:99-103, 2014
9. Danciu I, Cowan JD, Basford M, et al: Secondary use of clinical data: The Vanderbilt approach. J Biomed Inform 52:28-35, 2014
10. Landis JR, Koch GG: An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 33:363-374, 1977
11. Soga N, Yatabe Y, Kageyama T, et al: Review of bioptic Gleason scores by central pathologist modifies the risk classification in prostate cancer. Urol Int 95:452-456, 2015
12. D'Amico AV, Whittington R, Malkowicz SB, et al: Biochemical outcome after radical prostatectomy or external beam radiation therapy for patients with clinically localized prostate carcinoma in the prostate specific antigen era. Cancer 95:281-286, 2002
13. Mohler JL, Armstrong AJ, Bahnson RR, et al: Prostate cancer, version 1.2016. J Natl Compr Canc Netw 14:19-30, 2016

14. Brook RH, McGlynn EA, Cleary PD: Quality of health care: Part 2—Measuring quality of care. N Engl J Med 335:966-970, 1996
15. Penson DF: Assessing the quality of prostate cancer care. Curr Opin Urol 18:297-302, 2008
16. Neuman MD, Wirtalla C, Werner RM: Association between skilled nursing facility quality indicators and hospital readmissions. JAMA 312:1542-1551, 2014
17. Howlader N, Ries LAG, Mariotto AB, et al: Improved estimates of cancer-specific survival rates from population-based data. J Natl Cancer Inst 102:1584-1598, 2010
18. Chien AT, Rosenthal MB: Medicare's physician value-based payment modifier: Will the tectonic shift create waves? N Engl J Med 369:2076-2078, 2013
19. Burwell SM: Setting value-based payment goals: HHS efforts to improve U.S. health care. N Engl J Med 372:897-899, 2015
20. Centers for Medicare and Medicaid Services: 2016 Qualified Clinical Data Registries. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/Downloads/2016QCDRPosting.pdf
21. Peckham C: Medscape Physician Lifestyle Report. http://www.medscape.com/features/slideshow/lifestyle/2015/public/overview
22. Wei WQ, Teixeira PL, Mo H, et al: Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc. 23:e20-e27, 2015
23. Zhang S, Elhadad N: Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. J Biomed Inform 46:1088-1098, 2013
24. Warner JL, Levy MA, Neuss MN, et al: ReCAP: Feasibility and accuracy of extracting cancer stage information from narrative electronic health record data. J Oncol Pract 12:157-158, e169-e179, 2016
25. Kline RM, Bazell C, Smith E, et al: Centers for Medicare and Medicaid Services: Using an episode-based payment model to improve oncology care. J Oncol Pract 11:114-116, 2015
26. D'Amico AV, Renshaw AA, Cote K, et al: Impact of the percentage of positive prostate cores on prostate cancer-specific mortality for patients with low or favorable intermediate-risk disease. J Clin Oncol 22:3726-3732, 2004

APPENDIX

The steps taken to extract prostate-specific antigen (PSA) value and Gleason score are described in detail here; clinical T stage was extracted directly from a templated document as described in the pseudocode (Table 1).

Steps to get PSA value from narrative documents:

1a. From notes containing keyword "PSA," extract "PSA" and the immediate word following "PSA" to see how PSA scores were written in the notes:

array_combine(regexp_extract_all(notes_field, 'PSA\s*\w+'), '~')

1b. Subject matter experts convened to review the output of 1a and determined rules to exclude stop words and to include PSA values that may be separated from the "PSA" by several words or characters.

2. The following expression to capture PSA results into an array was selected for the analysis:

array_combine(regexp_extract_all (notes_field, 'PSA\s*(SCORE|VALUE|ROSE|RISING|REMAINS ELEVATED)?\s*(OF|ON|TO|WAS|IS|AT|\=)?\:?\s*([0-9]{1,4}/[0-9]{0,2}/?[0-9]{0,4})?(\s*[0-9]{1,2}\:[0-9]{1,2})?\s*([0-9]{1,3}\.?[0-9]*)|(undetectable|<\s*0.10)', 'i'), '~')

3. After splitting the above array into separate records, remove date time, date, and keywords from results to get the final PSA value.

4. Maximum PSA value in the defined period is chosen for the risk calculation.

Steps to get Gleason score:

1. Capture all clinical notes and pathology reports with the case-insensitive keyword "gleason."

2. Working with above captured notes, extract Gleason score into an array:

array_combine(regexp_extract_all(content,'[0-9]\s*\+\s*[0-9]\s*(\=\s*(10|[0-9]))?'), '~')

3. Split above array and parse score into primary and secondary scores.

4. Calculate total Gleason score by summing primary and secondary scores.