original report

Deep Learning for Natural Language Processing in Urology: State-of-the-Art Automated Extraction of Detailed Pathologic Prostate Cancer Data From Narratively Written Electronic Health Records

abstract

Purpose Entering all information from narrative documentation for clinical research into databases is time consuming, costly, and nearly impossible. Even high-volume databases do not cover all patient characteristics and drawn results may be limited. A new viable automated solution is machine learning based on deep neural networks applied to natural language processing (NLP), extracting detailed information from narratively written (eg, pathologic radical prostatectomy [RP]) electronic health records (EHRs).

Methods Within an RP pathologic database, 3,679 RP EHRs were randomly split into 70% training and 30% test data sets. Training EHRs were automatically annotated, providing a semiautomatically annotated corpus of narratively written pathologic reports with initially context-free gold standard encodings. Primary and secondary Gleason pattern, corresponding percentages, tumor stage, nodal stage, total volume, tumor volume and diameter, and surgical margin were variables of interest. Second, state-of-the-art NLP techniques were used to train an industry-standard language model for pathologic EHRs by transfer learning. Finally, accuracy of the named entity extractors was compared with the gold standard encodings.

Results Agreement rates (95% confidence interval) for primary and secondary Gleason patterns each were 91.3% (89.4 to 93.0), corresponding to the following: Gleason percentages, 70.5% (67.6 to 73.3) and 80.9% (78.4 to 83.3); tumor stage, 99.3% (98.6 to 99.7); nodal stage, 98.7% (97.8 to 99.3); total volume, 98.3% (97.3 to 99.0); tumor volume, 93.3% (91.6 to 94.8); maximum diameter, 96.3% (94.9 to 97.3); and surgical margin, 98.7% (97.8 to 99.3). Cumulative agreement was 91.3%.

Conclusion Our proposed NLP pipeline offers new abilities for precise and efficient data management from narrative documentation for clinical research. The scalable approach potentially allows the NLP pipeline to be generalized to other genitourinary EHRs, tumor entities, and other medical disciplines.

Clin Cancer Inform. © 2018 by American Society of Clinical Oncology

Sami-Ramzi Leyh-Bannurah
Zhe Tian
Pierre I. Karakiewicz
Ulrich Wolffgang
Guido Sauter
Margit Fisch
Dirk Pehrke
Hartwig Huland
Markus Graefen
Lars Budäus

Author affiliations and support information (if applicable) appear at the end of this article.

Corresponding author: Sami-Ramzi Leyh-Bannurah, MD, Martini-Klinik Prostate Cancer Center, Martinistrasse 52, 20246 Hamburg, Germany; e-mail: S.Bannurah@googlemail.com.

INTRODUCTION

One of the biggest challenges in clinical research is lack of either volume or detail in clinical research data. Even high-volume databases such as the SEER or National Inpatient Sample Database do not contain sufficient detail about the underlying disease such that most results drawn from these data sets might be limited. For example, SEER-related studies about metastatic prostate cancer (PCa) lack information on comorbidities, performance status, prostate-specific antigen response and type of androgen deprivation therapy, which is important for patient selection and prognosis.[1] Conversely, institutional or clinical trial data may provide far more detail in comparison. However, the relatively small patient

sample size makes conclusions drawn from these data sets difficult to generalize. Moreover, although the detailed information is generally present in the narrative clinical documentation, the extraction and generating of research data from such documentation are time-consuming, nonstandardized tasks that require human manpower and are costly. For example, the cost estimate of data management in a phase III trial with approximately 400 patients is well more than $1.2 million.[2] Data management requires many nurses and clinical research-associates to perform several rounds of quality control. Despite clinical data being the key to quality control and clinical outcome research, transferring narrative data computable for quantitative and qualitative analyses represents the main problem of either volume or detail in contemporary medical research.

With the advent of machine learning (ML) based on deep neural networks (DNN) applied to natural language processing (NLP), an automated approach may be a viable solution to generate detailed research data in high-volume capacity. NLP represents the intersection of computer science, linguistics, and machine intelligence.[3] Because language is probabilistic by nature, a corresponding probabilistic language model is required. For that reason, an ML approach is optimal for training of probabilistic language models. Moreover, special medical fields and tasks require special language models for information extraction.[4-7] Thus, NLP has great potential to automate the information-gathering process in the field of urology.

Another important benefit of such approach is that ML based on DNN represents a unified framework, which allows not only processing of text but of heterogeneous data types, such as images, audio, and any other type of data encoding.[8-10] Such an approach seems optimal to process an individual patient health file in conciliated manner. Specifically, an individual patient file consists of many different electronic health records (EHRs), such as imaging reports, surgical reports (eg, radical prostatectomy [RP]), and pathologic reports.[11]

Even a highly specific and standardized EHR such as a PCa pathology report is continuously revised and complemented (eg with genetic information).[12,13] Thus, previously reported, static, hand-crafted NLP rule sets to extract data from such reports, without supervised training of an ML model, represent an outdated approach that automatically suffers in performance over time.[14,15] Eventually, a mature system of data extraction would enable data to be built up systematically with far greater detail than he currently available quality for hypothesis testing, such as effective treatment of rare histologic subtypes of cancer. Moreover, with respect to open data, extracted data could be shared and merged more easily, particularly between institutions.

On the basis of these clinical considerations, we introduce an NLP pipeline to extract detailed pathologic PCa information from patients with have undergone RP. For that purpose, we relied on narratively written pathologic EHRs and a prospectively recorded institutional database as a reference standard that contains clinical characteristics, standardized information on RP specimens, lymph nodes, status, surgical margins, and follow-up data.

METHODS

All patients were surgically treated with RP in Martini-Clinic Prostate Cancer Center, Hamburg, Germany, between 2009 and 2017. All RP specimens were processed and examined by dedicated genitourinary pathologists of the Institute of Pathology, University Medical Center Hamburg-Eppendorf, Germany.[16] The pathologic EHR contained a macroscopic and microscopic description of the prostate specimen and lymph nodes with optional immunohistochemistry. The narratively written pathologic EHRs contained the following variables of interest (VOI): tumor stage (pT2a, pT2b, pT2c, pT3a, pT3b, pT4), primary and secondary Gleason patterns and corresponding percentages, surgical margin (R0, RX, R1), prostate specimen volume, tumor volume, und tumor maximum diameter and nodal stage (pN0, pNX, pN; Appendix Table A1). EHRs were entered into the institutional database by dedicated medical technical assistants and these served as the reference standard. All entered data were subsequently validated by either the treating surgeon or ward physician, serving as manual double entry of data.[17]

First, before NLP implementation, 3,679 RP EHRs overall were available as an a priori digital file and randomly split into 70% training (n = 2,575) and 30% test data sets (n = 1,104;

Appendix Table A2). Second, for each VOI, we followed a two-step approach. In a training data-generation step, we identified the part of the EHR that corresponded to the VOI we were trying to extract. For example, if a sentence contains a number from the gold encodings, which is surrounded by the keywords "tumor diameter" and "mm," such number within that sentence has a high likelihood of being the tumor diameter. We labeled the provided narrative EHRs with annotations based on subject experts, semiautomatically supported by using regular expressions, until the annotated EHRs were sufficient to serve as training data. In this step, we ensured that the entirety of the information we wanted to extract in the training data were captured (eg, 100% sensitivity).

Third, in a named entity recognition (NER) training step, an NER algorithm was trained on the basis of the annotated EHR specifically for the named entity we wanted to extract. The NER training step was done on the basis of the training data provided by the first step. Upon extraction, we need to distinguish between "tumor diameter is less than 10 mm" and " tumor diameter is 10 mm." Thus, the NER algorithm had to take into account the embedding of named entities in the sentence (ie, surrounding tokens or words). Our algorithm was based on a multilayered convolutional neural network, which was trained by back propagation.[18] Using the training data, we fine tuned parameters of the convolutional neural network until, to the best of our ability, accuracy no longer increased.

Finally, sensitivity analyses were performed to compare the accuracy of the NER to the gold standard encoding on a per-variable basis. Accuracy is the proportion of correctly classified positive reports among all PCa reports. Moreover, a mean accuracy was calculated cumulatively, based on all VOI. Statistical analyses were performed with the statistical package for R (R foundation for Statistical Computing, version 3.2.2; https://cran.r-project.org/bin/windows/base/old/3.2.2/).

RESULTS

Comparing NLP findings of 3,679 RP EHRs with the gold standard encodings, agreement rates (95% confidence interval) were as follows (Table 1), respectively: primary and secondary Gleason pattern, each 91.3% (89.4 to 93.0); corresponding Gleason percentages, 70.5% (67.6 to 73.3) and 80.9% (78.4 to 83.3); tumor stage, 99.3% (98.6 to 99.7); total RP specimen volume, 98.3% (97.3 to 99.0); tumor volume, 93.3% (91.6 to 94.8); tumor maximum diameter, 96.3% (94.9 to 97.3); surgical margin, 98.7% (97.8 to 99.3); and nodal stage, 98.7% (97.8 to 99.3). Specifically, the misclassification rates for registering different values versus missing the value were as follows (Appendix Table A3), respectively: primary and secondary Gleason pattern, each 8.6% versus 0.1%; corresponding Gleason percentages, 5.4% versus 24.1% and 8.6% versus 10.5%; tumor stage, 0.1% versus 0.6%; total RP specimen volume, 5.5% versus 3.9%; tumor volume, 1.0% versus 5.7%; tumor maximum diameter, 0.8% versus 3.0%; surgical margin, 0.6% versus 0.7%; and nodal stage, 0.2% versus 1.1%. The cumulative agreement rate was 91.3%.

DISCUSSION

We found that data acquired by the NLP configuration resulted in high agreement with the gold standard encodings. Such findings clearly demonstrate that NLP is applicable and sufficiently flexible to meet a highly specialized institutional need such as identifying specific VOI. Specifically, within this context, despite being narratively standardized, pathologic EHRs represent detailed medical data with a multitude of specific expressions and metrics. Taken together, the data extraction pipeline we propose can have a profound impact on PCa research by making more detailed data available for investigators. For example, already with the extraction of tumor volume, tumor maximum diameter, and Gleason proportions, clinically important questions with respect to tumor quantification can be addressed efficiently.

It is of note that the agreement rate for the primary Gleason proportion was < 90% (specifically, 70.5%). This low rate is based on the first version of the NLP pipeline, which was not robust with respect to combined Gleason patterns, specifically placement of primary and secondary Gleason patterns such as "Gleason 3+4=7" instead of "Gleason 4+3=7." As a solution, the NLP pipeline can be appended with a semantic reasoning module, which implements domain-specific data-cleansing capabilities and domain knowledge in a rule-based manner.

Table 1. Individual and Cumulative Agreement Rates of Pathologic Variables of Interest of the Natural Language Processing Generated Testing Database

	Digital Electronic Health Records (N = 1,104)			
Variable	Accuracy, %	95% CI	Wrong Value Assigned, %	No Value Assigned, %
Primary Gleason pattern	91	89 to 93	8.6	0.1
Primary Gleason pattern proportion, %	71	68 to 73	5.4	24
Secondary Gleason pattern	91	89 to 93	8.6	0.1
Secondary Gleason pattern proportion, %	81	78 to 83	8.6	11
Pathologic tumor stage	99	99 to 100	0.1	0.6
Prostate specimen volume, mL	98	97 to 99	5.5	3.9
Surgical margin status (R0 *v* RX *v* R1)	99	98 to 99	0.6	0.7
Tumor volume, mL	93	92 to 95	1.0	5.7
Tumor maximum diameter, mm	96	95 to 97	0.8	2.9
Nodal stage (pN0 *v* pNX *v* pN1)	99	98 to 99	0.2	1.1
Cumulative agreement	91			

NOTE. On the basis of the electronic health records in relation to the reference encodings of Martini-Clinic database.
Abbreviation: CI, confidence interval.

A strength of our study is the virtually completely validated, institutional data set, which represents a sufficiently large and comprehensive gold standard of high quality for the purpose of training and testing the NLP pipeline. Usually, medical gold standard encodings are limited for testing purposes. This is in marked contrast to nonmedical fields, where data are usually abundant, such as image and voice recognition.

Our cumulative agreement of 90% demonstrates that high performance can be achieved with the ML approach. This phenomenon was already observed in the field of image recognition, where, in the early 2000s, an elaborate rule-based approach was used and subsequently outperformed by DNN-based implementations.[19] Moreover, it has been demonstrated that such an approach manages to equal or outperform humans with respect to certain tasks.[20]

In marked contrast, historical series that reported about abstracting data from narrative written reports relied either on manually reviewed data as reference or had a limited number of VOI.[7] Moreover, due to the nature of the subject, recent technical advancements strongly outdate previous findings.[7] Specifically, mostly static, handcrafted NLP rule sets were used to extract data, without supervised training of an ML model.[14] Despite not using ML, it is of note that other studies achieved a good performance of analyzing EHRs.[11,15] These studies implemented workflows, which relied on comprehensive dictionaries of terms (eg, International Classification of Diseases, ninth revision, diagnosis).[11,15] Consequently, on the basis of these previous results and our own findings, we consider combining the high-level systematic approach of the International Classification of Diseases, ninth revision, ontology) with our flexible linguistics-based approach to have the greatest potential. Such a combination would likely outperform both isolated approaches, representing a consecutive chain of linguistics for data extraction, reasoning (eg, drug incompatibility), and decision-making (eg, risk classification, treatment recommendation).

On the basis of these considerations and the overall technical advancement in ML driven by DNN, there is a paradigm shift: The development of handcrafted feature generators for data extraction, for example (usually completely institutionally in a self-contained and customized way) is increasingly replaced by state-of-the-art NLP tools, which are freely accessible and theoretically can be trained for multiple purposes depending on the inputted data. In summary, it is not about developing a new algorithm any more but training existing DNN architectures with high-quality data such as the database we examined.

Outlooks include introducing new VOIs (eg, quantification of lymph node metastases burden where randomized control trials or population-based databases are not feasible), which would allow additional investigation of the question of imaging evasion and stratification according to oncologic risk profile. Such quantified burden is potentially superior to simple enumeration of positive lymph nodes, which have been highly investigated so far.[21]

Another prospect is the combination of pathologic EHRs with other medical EHRs (eg, from the fields of radiology or nuclear medicine). This would enable pathologic validation of imaging strategies for tumor burden thresholds that evade conventional morphologic criteria.

Finally, contrary to, for example, administrative codes or rigid preprogrammed search queries, parameters like medical notes, blood testing results, or imaging reports could be monitored prospectively. This would represent a true real-time approach to facilitate quality assurance and monitor for potential complications. Taken together, in systems with integrated EHRs, prospective surveillance could be extended to the outpatient setting or, as in our context, follow-up after primary treatment or patient-reported outcomes.[22-24]

On a technical level, NLP continuously advances in the fields of word representation, feature generators, neural network architecture, transfer learning, and the implementations for semantic reasoning. On a linguistic level, it is important to note that word meanings are encoded as word vectors in the presented approach, which enabled state-of-the-art synonym detection. These word vectors are provided as part of a pretrained language model (for each language). However, it does not contain medical terminology. Future work will focus on extending the vocabulary for adding word vectors by unsupervised training. Consequently, a much higher generalizability could be achieved (eg, tumor-free margin = R0 = negative resection border).

Nonetheless, our study has limitations. First, we used a specific and standardized German EHR, leading to homogeneous training and test data set. Generalizability in applying our NLP pipeline between different institutions is restricted by highly likely interinstitutional heterogeneity of narrative EHRs. Augmenting our training data with EHRs of other institutions and languages would mitigate this problem, because our NLP model would then generalize automatically. However, a reference standard similar to the Martini-Clinic database would be needed. Second, our EHRs are written in German. An existing pretrained German language model was used. In case of different languages, a corresponding language model would be required; these are freely available. Third, despite appearing self-explanatory, is important to note that the ML approach would need to be periodically recalibrated to account for changes in the reports in the event that new standards were established, narrative structures were changed, new guidelines were implemented, or simply that new author styles were introduced.

Fourth, our approach was only applied to aforementioned VOI. It has yet to be determined that our approach can be generalized to other and more complex named entities as well as categorization of whole paragraphs and documents. The latter would require integrating an additional document classifier into the NLP pipeline. For this task, a new text-classifier module or bidirectional long- and short-term memory recurrent neural network could be applied.

Fifth, at the current stage, we do not have 100% perfect extraction. A virtually perfect agreement still would be desirable with regard to the individual patient. However, all information that is extracted with high (> 90%) accuracy can be already be directly used in epidemiologic research. Because we would know the extent of the imperfection of the data a priori, many existing statistical methods could be used to account for missing data or data with measurement errors, to ensure we draw the appropriate conclusion using data obtained with NLP extraction methodology.[25]

Finally, a multidisciplinary team of clinical experts, and computer and data scientists participated in our study. Such a cross-disciplinary approach is essential but could represent a buy-in barrier for certain institutions.[6,26] However, it is important to acknowledge that a ground-breaking new algorithm does not have to be developed; open-source state-of-the-art algorithms were tailored for this specific medical task. In turn, limitations and drawbacks of closed-source, commercial alternatives such a licensing costs, vendor lock-in, and extensibility concerns do not apply.

In conclusion, we have demonstrated that contemporary, state-of-the-art NLP pipeline algorithms can successfully be applied in the field of urology to automatically extract highly specific variables of interest. If such an approach would be extended to more variables and other medical records across several institutions, analyses of multi-institutional NLP-generated databases could realistically provide a higher level of evidence compared with population-based databases.

DOI: https://doi.org/10.1200/CCI.18.00080
Published online on ascopubs.org/journal/cci on November 30, 2018.

AUTHOR CONTRIBUTIONS

Conception and design: Sami-Ramzi Leyh-Bannurah, Zhe Tian, Ulrich Wolffgang, Hartwig Huland, Markus Graefen, Lars Budäus
Administrative support: Sami-Ramzi Leyh-Bannurah, Hartwig Huland, Markus Graefen, Margit Fisch, Guido Sauter, Pierre I. Karakiewicz
Collection and assembly of data: Sami-Ramzi Leyh-Bannurah, Zhe Tian
Data analysis and interpretation: Sami-Ramzi Leyh-Bannurah, Zhe Tian
Manuscript writing: All authors
Final approval of manuscript: 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.

Sami-Ramzi Leyh-Bannurah
No relationship to disclose

Tian Zhe
Employment: IQVIA

Pierre I. Karakiewicz
No relationship to disclose

Ulrich Wolffgang
No relationship to disclose

Guido Sauter
No relationship to disclose

Margit Fisch
Consulting or Advisory Role: Boston Scientific (Inst)
Research Funding: Boston Scientific (Inst)
Travel, Accommodations, Expenses: Takeda

Dirk Pehrke
No relationship to disclose

Hartwig Huland
No relationship to disclose

Markus Graefen
No relationship to disclose

Lars Budäus
No relationship to disclose

Affiliations

Sami-Ramzi Leyh-Bannurah, **Dirk Pehrke**, **Hartwig Huland**, **Markus Graefen**, and **Lars Budäus**, Prostate Cancer Center Hamburg-Eppendorf; **Sami-Ramzi Leyh-Bannurah**, **Margit Fisch**, and **Guido Sauter**, University Medical Center Hamburg-Eppendorf, Hamburg; **Ulrich Wolffgang**, University of Muenster, Muenster, Germany; and **Zhe Tian** and **Pierre I. Karakiewicz**, University of Montreal Health Center, Montreal, Canada.

REFERENCES

1. Leyh-Bannurah SR, Gazdovich S, Budäus L, et al: Local therapy improves survival in metastatic prostate cancer. Eur Urol 72:118-124, 2017
2. Eisenstein EL, Lemons PW II, Tardiff BE, et al: Reducing the costs of phase III cardiovascular clinical trials. Am Heart J 149:482-488, 2005
3. Nadkarni PM, Ohno-Machado L, Chapman WW: Natural language processing: An introduction. J Am Med Inform Assoc 18:544-551, 2011
4. Gehrmann S, Dernoncourt F, Li Y, et al: Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLoS One 13:e0192360, 2018
5. Tian Z, Sun S, Eguale T, et al: Automated extraction of VTE events from narrative radiology reports in electronic health records: A validation study. Med Care 55:e73-e80, 2017
6. Wu JT, Dernoncourt F, Gehrmann S, et al: Behind the scenes: A medical natural language processing project. Int J Med Inform 112:68-73, 2018

7. Murff HJ, FitzHenry F, Matheny ME, et al: Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 306:848-855, 2011
8. Chiarelli PA, Hauptman JS, Browd SR: Machine learning and the prediction of hydrocephalus: Can quantitative image analysis assist the clinician? JAMA Pediatr 172:116-118, 2018
9. Gulshan V, Peng L, Coram M, et al: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316:2402-2410, 2016
10. Blum ES, Porras AR, Biggs E, et al: Early detection of ureteropelvic junction obstruction using signal analysis and machine learning: A dynamic solution to a dynamic problem. J Urol 199:847-852, 2017
11. Nead KT, Gaskin G, Chester C, et al: Androgen deprivation therapy and future Alzheimer's disease risk. J Clin Oncol 34:566-571, 2016
12. Epstein JI, Allsbrook WC Jr, Amin MB, et al: The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. Am J Surg Pathol 29:1228-1242, 2005
13. Epstein JI, Egevad L, Amin MB, et al: The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of grading patterns and proposal for a new grading system. Am J Surg Pathol 40:244-252, 2016
14. 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
15. Gregg JR, Lang M, Wang LL, et al.: Automating the determination of prostate cancer risk strata from electronic medical records. JCO Clin Cancer Inform 10.1200/CCI.16.00045
16. Sauter G, Steurer S, Clauditz TS, et al: Clinical utility of quantitative Gleason grading in prostate biopsies and prostatectomy specimens. Eur Urol 69:592-598, 2015
17. Paulsen A, Overgaard S, Lauritsen JM: Quality of data entry using single entry, double entry and automated forms processing--An example based on a study of patient-reported outcomes. PLoS One 7:e35087, 2012
18. Rumelhart, DE, Hinton, GE, Williams, RJ: Learning representations by back-propagating errors. Nature 323:533-536, 1986
19. Krizhevsky A, Sutskever I, Hinton GE: Imagenet classification with deep convolutional neural networks.Presented at Advances in Neural Information Processing Systems, Lake Tahoe, NV, December 3-6, 2012
20. Esteva A, Kuprel B, Novoa RA, et al: Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115-118, 2017
21. Gandaglia G, Fossati N, Zaffuto E, et al: Development and internal validation of a novel model to identify the candidates for extended pelvic lymph node dissection in prostate cancer. Eur Urol 72:632-640, 2017
22. Pompe RS, Tian Z, Preisser F, et al: Short- and long-term functional outcomes and quality of life after radical prostatectomy: Patient-reported outcomes from a tertiary high-volume center. Eur Urol Focus 3:615-620, 2017
23. Adam M, Tennstedt P, Lanwehr D, et al: Functional outcomes and quality of life after radical prostatectomy only versus a combination of prostatectomy with radiation and hormonal therapy. Eur Urol 71:330-336, 2017
24. Martin NE, Massey L, Stowell C, et al: Defining a standard set of patient-centered outcomes for men with localized prostate cancer. Eur Urol 67:460-467, 2015
25. Hu C, Lin D: Cox regression with covariate measurement error. Scand J Stat 29:637-655, 2002
26. Hashimoto DA, Rosman G, Rus D, et al: Artificial intelligence in surgery: Promises and perils. Ann Surg 268:70-76, 2018

Appendix

Table A1. Proportions of Categorical Pathologic Prostate Cancer Variables of Interest of 3,679 Patients With Prostate Cancer Who Were Treated With Radical Prostatectomy

Variable	No.	%
Primary Gleason pattern		
3	2,632	72
4	887	24
5	160	4.4
Secondary Gleason pattern		
3	914	25
4	2,528	69
5	237	6.4
Pathologic tumor stage		
pT2a	243	6.6
pT2b	5	0.1
pT2c	2,024	55
pT3a	832	23
pT3b	559	15
pT4	16	0.4
Surgical margin status		
R0	2,910	79
R1	741	20
RX	18	0.5
Nodal stage		
N0	2,801	76
N1	504	14
NX	374	10

NOTE. On the basis of the reference encodings of Martini-Clinic database.

Table A2. Pseudocode Describing the Natural Language Processing Pipeline

Step	Pseudocode
1	Provisioning of 3,679 pathologic electronic health records (PDF format). Random splitting of the data into 70% training data and 30% validation data
2	Using Martini Clinic gold standard encodings and regular expressions for semiautomated generation of training data
3	Training of spaCy* named entity recognition by training data from step 2. The training is based on back propagation.
4	Testing of the spaCy named entity recognition: using the remaining 30% validation data and testing against the Martini Clinic gold standard encodings

*https://spacy.io/.

Table A3. Discrepancy Analysis of Pathologic Variables of Interest of the Natural Language Processing–Generated Testing Database

	Digital Electronic Health Records (N = 1,104)		
Variable of Interest	Wrong Value Assigned, %	No Value Assigned, %	Discrepancy Notes
Primary Gleason pattern	8.6	0.1	Gleason pattern contained a decimal (eg, Gleason 3.0+4)
Primary Gleason pattern proportion, %	5.4	24	Gleason pattern proportion was sometimes given as percentage (80%), sometimes exclusively as absolute number in mL (15%), sometimes missing (5%)
Secondary Gleason pattern	8.6	0.1	Gleason pattern contained a decimal (eg, Gleason 3.0+4); sometimes in Gleason pattern 3+3, 4+4, or 5+5; the second pattern value was not entered
Secondary Gleason pattern proportion, %	8.6	11	Gleason pattern proportion was sometimes given as percentage (80%), sometimes as absolute number in mL (15%), sometimes missing (5%). Sometimes in Gleason pattern 3+3, 4+4, or 5+5; the second pattern proportion value was not entered
Pathologic tumor stage	0.1	0.6	Minor typing error within database (eg, "p3" instead of "pT3")
Radical prostatectomy specimen total volume, mL	5.5	3.9	mg instead of milliliter or gram
Surgical margin status (R0 *v* RX *v* R1)	0.6	0.7	Sometimes no final surgical margin assessment was given, but was stated as artifact. In this case, the information "RX" was missing.
Radical prostatectomy specimen tumor volume, mL	1.0	5.7	(a) "mg" instead of "mL" (b) If Gleason pattern proportion was not entered either as percentage or as mL, the RP specimen total volume was missing in most cases.
Radical prostatectomy specimen tumor maximum diameter, mm	0.8	2.9	Information was missing, particularly when RP specimen tumor volume was given.
Nodal stage (pN0 *v* pNX *v* pN1)	0.2	1.1	Minor typing error within database (eg, "0" instead of "pN0")

NOTE. On the basis of the electronic health records in relation to the reference encodings of Martini-Clinic database.
Abbreviation: RP, radical prostatectomy.