Abstract

**IMPORTANCE** Although tissue-based gene expression testing has become widely used for prostate cancer risk stratification, its prognostic performance in the setting of clinical care is not well understood.

**OBJECTIVE** To develop a linkage between a prostate genomic classifier (GC) and clinical data across payers and sites of care in the US.

**DESIGN, SETTING, AND PARTICIPANTS** In this cohort study, clinical and transcriptomic data from clinical use of a prostate GC between 2016 and 2022 were linked with data aggregated from insurance claims, pharmacy records, and electronic health record (EHR) data. Participants were anonymously linked between datasets by deterministic methods through a deidentification engine using encrypted tokens. Algorithms were developed and refined for identifying prostate cancer diagnoses, treatment timing, and clinical outcomes using diagnosis codes, Common Procedural Terminology codes, pharmacy codes, Systematized Medical Nomenclature for Medicine clinical terms, and unstructured text in the EHR. Data analysis was performed from January 2023 to January 2024.

**EXPOSURE** Diagnosis of prostate cancer.

**MAIN OUTCOMES AND MEASURES** The primary outcomes were biochemical recurrence and development of prostate cancer metastases after diagnosis or radical prostatectomy (RP). The sensitivity of the linkage and identification algorithms for clinical and administrative data were calculated relative to clinical and pathological information obtained during the GC testing process as the reference standard.

**RESULTS** A total of 92,976 of 95,578 (97.2%) participants who underwent prostate GC testing were successfully linked to administrative and clinical data, including 53,871 who underwent biopsy testing and 39,105 who underwent RP testing. The median (IQR) age at GC testing was 66.4 (61.0-71.0) years. The sensitivity of the EHR linkage data for prostate cancer diagnoses was 85.0% (95% CI, 84.7%-85.2%), including 80.8% (95% CI, 80.4%-81.1%) for biopsy-tested participants and 90.8% (95% CI, 90.5%-91.0%) for RP-tested participants. Year of treatment was concordant in 97.9% (95% CI, 97.7%-98.1%) of those undergoing GC testing at RP, and 86.0% (95% CI, 85.6%-86.4%) among participants undergoing biopsy testing. The sensitivity of the linkage was 48.6% (95% CI, 48.1%-49.1%) for identifying RP and 50.1% (95% CI, 49.7%-50.5%) for identifying prostate biopsy.

**CONCLUSIONS AND RELEVANCE** This study established a national-scale linkage of transcriptomic and longitudinal clinical data yielding high accuracy for identifying key clinical junctures, including (continued)
Abstract (continued)

diagnosis, treatment, and early cancer outcome. This resource can be leveraged to enhance understandings of disease biology, patterns of care, and treatment effectiveness.


Introduction

Within the past decade, there have been major increases in the scale and personalization of prostate cancer data. Prognostic tissue-based gene expression (genomic) tests have become widely used in clinical decision-making at multiple junctures in the disease. In parallel, there have been marked increases in the volume and detail of clinical-practice data (ie, those generated in routine clinical care) through the rapid adoption of electronic health record (EHR) systems, digital health tools, and broader access to insurance claims. Among genomic tests, the Decipher (Veracyte, Inc) prostate genomic classifier (GC), a 22-gene expression signature, has been consistently identified as a prognostic measure for key clinical end points, including adverse pathology, biochemical recurrence (BCR), metastasis, and death from prostate cancer. Clinical-practice data capturing contemporaneous and nationally representative prostate cancer care have been increasingly used to explore treatment patterns and make inferences about treatment effectiveness. These separate but supplementary data sources offer unrealized opportunities if they can be aligned. Clinical-practice data analyses frequently lack clinical and biologic detail, are prone to unmeasured confounder effects, and often fail to agree with evidence from randomized trials. Genomic analyses, by virtue of their novelty, have lacked the opportunity to assess long-term end points or account for clinical awareness of test results, their impact on clinical decisions, and changing treatment landscapes. Thus, integration of clinical and genomic data sources may generate deeper understanding of disease management than may be possible by independent analyses.

Despite their potential to enhance insights about disease biology and treatment effectiveness, to our knowledge, no comprehensive longitudinal linkages of clinical and transcriptomic data have yet been developed for prostate cancer. Recent linkages of the Surveillance Epidemiology and End Results database with GC results have characterized patterns of use in a nationally representative subset, but have not evaluated long-term cancer outcomes such as recurrence or metastasis. Linkages in other cancer types have integrated only limited genomic data within existing clinical cancer registries, generating new information about prognosis and cost-effectiveness. By comparison, alignment of clinical and genomic data from the GC platform is especially promising because the latter includes not only the validated GC reported to ordering practitioners but also complete transcriptomic profiling of prostate tumors, including over 1.4 million features including more than 46,000 coding and noncoding genes, as well as hundreds of locked gene expression signatures, a scale that has not previously been achieved in prostate cancer research. If fully realized, this resource may be leveraged to refine personalized estimates of cancer outcome and treatment selection.

To evaluate the prognostic value of the GC in the clinical practice setting and inform deeper explorations of disease outcome at national scale, we developed a novel linkage between the GC and clinical practice patient data in the US across payers and sites of care. In this report, we describe the data linkage structure and validate and refine algorithms for identifying key prostate cancer events, including dates of diagnosis, treatment, BCR, and metastasis. This linkage serves as the foundation for future analyses of gene expression and clinical outcomes in prostate cancer.
Methods

The use of deidentified data for this cohort study was deemed non–human participants research by the Yale University institutional review board; thus, the need for informed consent was waived, in accordance with 45 CFR §46. This analysis was conducted in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines (eAppendix 1 in Supplement 1).21

Study Objective

The primary objective of this study was to develop a comprehensive linkage between prostate cancer transcriptomic and clinical and administrative data. The secondary objectives were to refine algorithms for identifying cancer diagnosis, treatment, and outcome events in clinical-practice data and examine their accuracy.

Data Sources and Linkage Structure

The first data source consisted of clinical and transcriptomic data from GC testing obtained during clinical use from prostate biopsy and radical prostatectomy (RP) testing between 2016 and 2022. As part of the ordering process, clinicians submitted a test requisition with clinical and pathological information, a pathology report, and tissue sample that was then verified to contain prostate cancer suitable for tumor testing. The GC score, an expression signature derived from 22 features associated with prostate cancer outcome and scaled from 0.0 to 1.0, was calculated for all specimens using previously described methods.5,22 Participant data were deidentified from clinical use in accordance with the Safe Harbor method described in the Health Insurance Portability and Accountability Act Privacy Rule 45 CFR 164.514(b) and (c) and were included in the prospectively collected Genomics Resource for Intelligent Discovery registry.23 The second data source included insurance claims, pharmacy records, and EHR data (Clarivate). Participants were anonymously and securely linked between datasets by deterministic methods through a deidentification engine using tokenization (Datavant) (Figure 1). This process involves assigning unique anonymized tokens generated using cryptographic algorithms substituting identifiers (first and last name, gender, and date of birth) for...
exchange between the data sources. The linkage algorithm matched records in 2 steps, requiring agreement in name, gender, and date of birth (Figure 1).

Data Elements

Linked data consisted of structured EHR data, including laboratory data files and unstructured EHR text strings. Administrative claims codes were compiled, including Common Procedural Terminology codes, pharmacy codes, and Systematized Medical Nomenclature for Medicine clinical terms (eTable 1 in Supplement 1). Participant-level genomic data included GC scores (range, 0.0-1.0; low, <0.45; intermediate, 0.45-0.60; and high, >0.60) and Decipher Genomics Resource for Intelligent Discovery whole transcriptome profiles conducted on each tumor specimen. Clinical data were compiled including prostate-specific antigen (PSA) level at diagnosis or RP, and risk groupings (Cancer of the Prostate Risk Assessment, and postsurgical Cancer of the Prostate Risk Assessment).

Patient Selection

We required that eligible participants achieve a valid linkage between EHR and genomic data sources as assessed through a perfect match of token 1 and token 2, which encrypted patients’ name, gender, and date of birth. The linked participants were stratified into mutually exclusive biopsy-tested and RP-tested groups on the basis of the sample type of the first GC test performed.

Outcomes

We assessed the incidence of other prostate cancer events not captured in GC test requisitions between January 1, 2021, to March 1, 2022 (eTable 2 in Supplement 1). For these events, we adapted a claims-based approach described by Freedland et al60 and calculated the proportion of participants identified by administrative claims codes, EHR text, or both methods (eTable 3 in Supplement 1). The clinical events included (1) prostate cancer diagnosis, (2) prostate biopsy, (3) RP, (4) BCR, (5) any metastasis, and (6) prostate cancer distant metastasis. The detailed methods for identifying the clinical events are presented in eAppendix 1 in Supplement 1.

Statistical Analysis

Data analysis was performed from January 2023 to January 2024. We described the proportion of participants successfully linked between EHR and genomic data sources. Within the broader sample, we characterized clinical, pathologic, and genomic characteristics. Next, we evaluated the agreement between the identified events from EHR data and that verified through the GC testing process. These included (1) sensitivity for detection of prostate cancer diagnoses in the linkage, (2) concordance between year of prostate biopsy and RP, and (3) incidence of BCR and metastasis (eAppendix 1 in Supplement 1).

We conducted the following prespecified sensitivity analyses. First, we examined differential use of the metastasis definition and established agreement based on clinical and pathological variables, including National Comprehensive Cancer Network risk group. Second, we evaluated temporal associations between PSA values and clinical events as identified in the EHR data. For this analysis, we selected participants with 1 or more PSA test result and calculated the logarithm of the mean PSA within 30-day time periods relative to the events of interest (RP and metastasis). In the event of multiple PSA tests within a given period, the mean value was used. We used Mann-Whitney U and Wilcoxon signed-rank tests to compare the distribution of mean PSA values between time periods.

Missing data were handled by pairwise detection. For each analysis, participants were selected according to the presence of required variables for calculating sensitivity, concordance, or event incidence. The analyses were conducted using R statistical software version 4.3.0 (R Project for Statistical Computing) with extension packages ggplot2 version 3.4.2 and table1 version 1.4.3. The
statistical tests were 2 sided, and an α level of .05 was prespecified to determine statistical significance.

Results

Sample Characteristics
A total of 92,976 of 95,578 participants (97.2%) with GC prostate testing from 2016 to 2022 were successfully linked to EHR data, including 53,871 from biopsy and 39,105 from RP testing (Table 1). There were 5908 participants who had multiple GC tests. The median (IQR) age at GC testing was 66.4 (61.0-71.0) years. Biopsy-tested individuals were older than those tested at RP (median [IQR] age, 68 [62-73] years vs 65 [60-69] years). Gleason grade group data were complete for 99.9% of

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Biopsy tested (n = 53,871)</th>
<th>RP tested (n = 39,105)</th>
<th>Overall (N = 92,976)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at testing, y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>67.4 (7.86)</td>
<td>64.3 (6.95)</td>
<td>66.1 (7.64)</td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>68.0 (62.0-73.0)</td>
<td>65.0 (60.0-69.0)</td>
<td>66.4 (61.0-71.0)</td>
</tr>
<tr>
<td>Missing</td>
<td>2415 (4.5)</td>
<td>30 (0.1)</td>
<td>2445 (2.6)</td>
</tr>
<tr>
<td>Gleason grade group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>19,347 (35.9)</td>
<td>1859 (4.8)</td>
<td>21,206 (22.8)</td>
</tr>
<tr>
<td>2</td>
<td>21,952 (40.7)</td>
<td>16,315 (41.7)</td>
<td>38,267 (41.2)</td>
</tr>
<tr>
<td>3</td>
<td>8,113 (15.1)</td>
<td>11,966 (30.6)</td>
<td>20,079 (21.6)</td>
</tr>
<tr>
<td>4</td>
<td>2,778 (5.2)</td>
<td>3,225 (8.2)</td>
<td>6,003 (6.5)</td>
</tr>
<tr>
<td>5</td>
<td>1,666 (3.1)</td>
<td>5,696 (14.6)</td>
<td>7,362 (7.9)</td>
</tr>
<tr>
<td>Missing</td>
<td>15 (&lt;0.1)</td>
<td>44 (0.1)</td>
<td>59 (0.1)</td>
</tr>
<tr>
<td>Most recent pretreatment prostate-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>specific antigen level, ng/mL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>8.86 (23.7)</td>
<td>7.92 (11.6)</td>
<td>8.60 (21.0)</td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>6.20 (4.72-8.72)</td>
<td>6.00 (4.02-9.30)</td>
<td>6.19 (4.60-8.90)</td>
</tr>
<tr>
<td>Missing</td>
<td>1699 (3.2)</td>
<td>19,364 (49.5)</td>
<td>21,063 (22.7)</td>
</tr>
<tr>
<td>CAPRA (biopsy) or CAPRA-S (RP) category</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>19,849 (36.8)</td>
<td>3,431 (8.8)</td>
<td>NA</td>
</tr>
<tr>
<td>Intermediate</td>
<td>24,420 (45.3)</td>
<td>10,113 (25.9)</td>
<td>NA</td>
</tr>
<tr>
<td>High</td>
<td>5,896 (10.9)</td>
<td>6,064 (15.5)</td>
<td>NA</td>
</tr>
<tr>
<td>Missing</td>
<td>3,706 (6.9)</td>
<td>19,497 (49.9)</td>
<td>NA</td>
</tr>
<tr>
<td>Pathologic feature</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seminal vesicle invasion</td>
<td>NA</td>
<td>7,662 (19.6)</td>
<td>NA</td>
</tr>
<tr>
<td>Lymph node involvement</td>
<td>NA</td>
<td>1,762 (4.5)</td>
<td>NA</td>
</tr>
<tr>
<td>Extraprostatic extension</td>
<td>NA</td>
<td>22,364 (57.2)</td>
<td>NA</td>
</tr>
<tr>
<td>Surgical margin positive</td>
<td>NA</td>
<td>20,754 (53.1)</td>
<td>NA</td>
</tr>
<tr>
<td>Biopsy cores involved, %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>0.32 (0.25)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>0.25 (0.14-0.50)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Missing</td>
<td>1 (&lt;0.1)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Genomic classifier category</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low (&lt;0.45)</td>
<td>27,661 (51.3)</td>
<td>15,113 (38.6)</td>
<td>42,774 (46.0)</td>
</tr>
<tr>
<td>Intermediate (0.45-0.60)</td>
<td>9,815 (18.2)</td>
<td>6,616 (16.9)</td>
<td>16,431 (17.7)</td>
</tr>
<tr>
<td>High (&gt;0.60)</td>
<td>16,395 (30.4)</td>
<td>17,376 (44.4)</td>
<td>33,771 (36.3)</td>
</tr>
<tr>
<td>Genomic classifier score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>0.48 (0.25)</td>
<td>0.56 (0.27)</td>
<td>0.51 (0.26)</td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>0.43 (0.27-0.67)</td>
<td>0.55 (0.33-0.80)</td>
<td>0.48 (0.29-0.73)</td>
</tr>
</tbody>
</table>

Abbreviations: CAPRA, Cancer of the Prostate Risk Assessment; CAPRA-S, postsurgical Cancer of the Prostate Risk Assessment; NA, not applicable; RP, radical prostatectomy.

SI conversion factor: To convert prostate-specific antigen to micrograms per liter, multiply by 1.
patients. Participants undergoing biopsy testing more commonly had Gleason grade group 1 (19,347 participants [35.9%]) and 2 (21,952 participants [40.7%]) tumors compared with participants undergoing testing from prostatectomy specimens (1,859 participants [4.8%] with Gleason grade group 1 tumors and 16,315 participants [41.7%] with Gleason grade group 2 tumors). GC scores were lower in biopsy vs RP tests (median [IQR] GC score, 0.43 [0.27-0.67] vs 0.55 [0.33-0.80]).

Sensitivity of Linkage for Prostate Cancer Diagnosis, Biopsy, and RP
The sensitivity, concordance, and incidence of prostate cancer outcomes are summarized in Table 2. The overall sensitivity of linkage for identifying prostate cancer diagnosis was 85.0% (78,989 of 92,976 participants; 95% CI, 84.7%-85.2%). The sensitivity of linkage for prostate cancer diagnosis was 80.8% (43,501 of 53,871 participants; 95% CI, 80.4%-81.1%) among biopsy-tested individuals and 90.8% (35,488 of 39,105 participants; 95% CI, 90.5%-91.0%) among RP-tested individuals. Prostate cancer diagnoses was identified using administrative claims codes in 67,523 cases (85.5%; 95% CI, 85.2%-85.7%), both codes and EHR text in 11,257 cases (14.3%; 95% CI, 14.0%-14.5%), and text only in 209 cases (0.26%; 95% CI, 0.23%-0.30%).

The sensitivity of the linkage for identification of RP was 48.6% (19,008 of 39,105 participants; 95% CI, 48.1%-49.1%) among those who received GC testing from their RP sample. The sensitivity for identifying biopsy was 50.1% (26,990 of 53,871 participants; 95% CI, 49.7%-50.5%) among those who received GC testing from a prostate biopsy sample. The year of prostate biopsy was concordant between GC test requisitions and EHR data in 86.0% of cases (23,220 of 26,990 participants; 95% CI, 85.6%-86.4%). Similarly, year of RP was concordant between data sources in 97.9% of cases (18,604 of 19,008 participants; 95% CI, 97.7%-98.1%). In 0.26% of cases (49 participants; 95% CI, 0.20%-0.34%), RP was identified earlier in EHR data, whereas in 1.9% of cases (355 participants; 95% CI, 1.7%-2.1%), dates of RP were identified to be later in EHR data compared with GC test requisitions (Figure 2).

Incidence of BCR and Metastasis
Among 92,976 linked participants, 5,895 (6.3%; 95% CI, 6.2%-6.5%) had evidence of BCR, including 1,711 (3.2%; 95% CI, 3.0%-3.3%) who underwent biopsy and 4,184 (10.7%; 95% CI, 10.4%-11.0%) who underwent RP GC testing (Table 2). Most participants with BCR events were identified on the basis of diagnosis code (96.3%; 95% CI, 95.8%-96.8%), with 0.05% (95% CI, 0.02%-0.15%) additionally

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Biopsy tested (n = 53,871)</th>
<th>RP tested (n = 39,105)</th>
<th>Overall (N = 92,976)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity Identifying prostate cancer diagnosis</td>
<td>43,501/53,871 (80.8%) [80.4-81.1]</td>
<td>35,488/39,105 (90.8%) [90.5-91.0]</td>
<td>78,989/92,976 (85.0%) [84.7-85.2]</td>
</tr>
<tr>
<td>Identifying RP procedure</td>
<td>26,990/53,871 (50.1%) [49.7-50.5]</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Identifying RP</td>
<td>NA</td>
<td>19,008/39,105 (48.6%) [48.1-49.1]</td>
<td>NA</td>
</tr>
<tr>
<td>Concordance Between year of prostate biopsy and year of testing</td>
<td>23,220/26,990 (86.0%) [85.6-86.4]^a</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Between year of RP and year of testing</td>
<td>NA</td>
<td>18,604/19,008 (97.9%) [97.7-98.1]^a</td>
<td>NA</td>
</tr>
<tr>
<td>Incidence Biochemical recurrence diagnosis</td>
<td>1,711/53,871 (3.2%) [3.0-3.3]</td>
<td>4184/39,105 (10.7%) [10.4-11.0]</td>
<td>5,895/92,976 (6.3%) [6.2-6.5]</td>
</tr>
<tr>
<td>Metastasis diagnosis</td>
<td>2,744/53,871 (5.1%) [5.0-5.3]</td>
<td>4500/39,105 (11.5%) [11.2-11.8]</td>
<td>7,244/92,976 (7.8%) [7.6-8.0]</td>
</tr>
<tr>
<td>Prostate cancer-specific metastasis in restricted subpopulation^c</td>
<td>361/33,379 (1.1%) [1.0-1.2]</td>
<td>931/25,556 (3.6%) [3.4-3.9]</td>
<td>1,292/58,395 (2.2%) [2.1-2.3]</td>
</tr>
</tbody>
</table>

Abbreviations: NA, not applicable; RP, radical prostatectomy.

^a Restricted to biopsy-tested participants with evidence of prostate biopsy procedure from claims or electronic health record data (26,990 of a total 53,871 biopsy tested).

^b Restricted to RP-tested participants with evidence of RP from linked claims or electronic health record data (19,008 of a total 39,105 RP tested).

^c Restricted to participants with no other cancer diagnoses except nonmelanoma skin cancer and no evidence of de novo prostate metastasis.
identified though unstructured text alone, and 3.7% (95% CI, 3.2%-4.2%) identified through both text and codes.

A total of 7244 participants (7.8%; 95% CI, 7.6%-8.0%) had evidence of metastasis, including 2744 (5.1%; 95% CI, 5.0%-5.3%) who underwent testing from biopsy samples and 4500 (11.5%; 95% CI, 11.2%-11.8%) who underwent testing from RP samples. Metastasis was identified by diagnosis code only in 6997 cases (96.6%; 95% CI, 96.1%-97.0%), by code and EHR text in 225 cases (3.1%; 95% CI, 2.7%-3.5%), and by EHR text alone in 22 cases (0.3%; 95% CI, 0.2%-0.5%).

When considering 58,935 individuals with no other cancer diagnoses except nonmelanoma skin cancer and no evidence of de novo prostate metastasis, 1292 (2.2%; 95% CI, 2.1%-2.3%) had evidence of distant prostate cancer metastasis, including 361 of 33,379 (1.1%; 95% CI, 1.0%-1.2%) biopsy-tested individuals and 931 of 25,556 (3.6%; 95% CI, 3.4%-3.9%) RP-tested patients (Table 2). Most prostate cancer distant metastasis events (98.0%; 95% CI, 97.1%-98.6%) were identified according to diagnosis codes, with 0.2% (95% CI, 0.1%-0.7%) identified through unstructured text, and 1.8% (95% CI, 1.2%-2.7%) identified through a combination of text and diagnosis codes.

Sensitivity Analyses
Results of sensitivity analyses by National Comprehensive Cancer Network risk group showed increasing incidence of metastases in higher risk groups (eTable 4 in Supplement 1). Among the subset with PSA values and evidence of RP, median (IQR) PSA levels decreased from 5.5 (3.8-9.8)
ng/mL (to convert to micrograms per liter, multiply by 1) in the 30- to 60-day period before treatment to 0.03 (0.01-0.10) ng/mL in the full 30- to 60-day period after treatment. PSA levels similarly decreased from a median (IQR) of 3.48 (0.23-7.52) ng/mL in the 30- to 60-day window before metastatic diagnosis to 0.43 (0.08-2.53) ng/mL in the 30- to 60-day period following after diagnosis (Figure 3). Additional results are shown in eAppendix 2 in Supplement 1.

Discussion

In this cohort study, we established a national linkage of prostate cancer transcriptomic and longitudinal clinical-practice data and found high accuracy for identifying key clinical junctures, including diagnosis, treatment, and cancer outcome. By virtue of its comprehensive scale and level of genomic detail—including a widely validated genomic signature and full transcriptome analysis for all participants—this linkage can be leveraged to enhance the understanding of determinants of cancer outcome, treatment effectiveness, and patterns of care delivery. Furthermore, by connecting whole transcriptomic profiling of more than 90,000 prostate cancers diagnosed in the contemporary era with clinicopathologic, treatment, and outcome data across payers and sites of care, we can further refine our understanding of these complex biological and clinical processes.

Figure 3. Distribution of Prostate-Specific Antigen (PSA) Values in Relation to Index Clinical Event (Radical Prostatectomy and Metastasis) Assessed Using Claims, Pharmacy Records, and Electronic Health Record Data

Median PSA values were assessed in the 30- to 60-day periods before (orange shading) and after (blue shading) event. Each point represents the log of the mean PSA value of a patient in the time window. The error bars denote IQRs of log mean PSA value. To convert PSA to micrograms per liter, multiply by 1.
care delivery, this resource can enhance diversity and representation in translational prostate cancer research. Accurate assessment of clinical events is necessary to conduct meaningful analyses from large-scale genomic repositories. To this end, we have not only defined the nature of this clinical-transcriptomic linkage but have also defined and analyzed algorithms to examine key clinical junctures relating to diagnosis, treatment, recurrence, and metastasis.

To our knowledge, this is the largest linkage of longitudinal clinical and genomic data to date in prostate cancer. We successfully linked 97.2% of participants with clinical data obtained from several sources, including insurance claims across payers, pharmacy records, and EHR data obtained in the clinical-practice setting. Furthermore, by including the majority of participants undergoing GC testing in the clinical-practice setting, the participants included in this linkage are demographically representative of those tested in the contemporary era. This work can complement findings from other large transcriptome-wide association studies that have also identified potential prostate cancer susceptibility loci, but lack detail about treatment exposure and clinical outcome.31,32

Among participants with identified EHR data evidence of RP and biopsy, clinical-practice data-based algorithms reliably identified treatment timing and sequence, achieving 97.9% concordance for the timing of RP and 86.0% concordance for diagnosis among participants who had GC calculated from a biopsy specimen. Moreover, through pharmacy records and claims, we were able to quantify the duration of systemic therapy and account for time-dependent exposures to systemic therapy, which can be leveraged for larger-scale pharmacogenomic study in prostate cancer in the future.33 However, another notable finding from this work is that large gaps in clinical data will exist for a subset of linked participants, including those without evidence of clinical or claims-based activity, thus increasing risks of misascertainment. Such gaps are likely to explain the modest sensitivity for detecting RP (48.6%) and prostate biopsy (50.1%). These considerations underscore the need for careful selection based on evidence of continuity within claims.

Through this work we refined algorithms for identifying metastasis after prostate cancer diagnosis and treatment. Using a hierarchical definition of metastasis incorporating diagnosis codes, text, and treatments for distant metastasis after diagnosis or treatment for localized disease, we build on prior algorithms for identifying these events in administrative claims.30,34 These procedures are notable for several reasons. First, we apply these algorithms from clinical and administrative data aggregated from multiple data sources rather than a database from a single payer. Second, we separately validate these approaches for participants undergoing biopsy and RP testing, a distinction that is relevant given potential differences in the time from initial diagnosis and manner of treatment. Furthermore, as a novel contribution, we incorporate laboratory PSA values to show that these are temporally related to the first incidence of a metastasis diagnosis and treatment initiation. These approaches can be incorporated into other prostate cancer studies using clinical-practice data across multiple payers and EHR sources.

This work is the first, to our knowledge, to connect transcriptomic profiling and longitudinal clinical-practice data at a comprehensive, national scale. As a result, future directions from this research include prognostic evaluation of GC testing in a clinical-practice setting. In addition, this resource can be used to develop or refine genomic signatures associated with prostate cancer outcome and treatment response. By virtue of its scale, there are also opportunities to extend the study of clinical scenarios with greater uncertainty, such as the prognostic impact of genomic classifiers on the outcome of patients managed with active surveillance, as well as the impact of testing results on clinical management decisions.35 Furthermore, this data resource can also enhance future effectiveness research by providing a more complete representation of treatment sequence, chronicity, and baseline genomic profile, which has not been examined.

Limitations

This study has limitations that should be mentioned. Because most clinical events were identified through administrative claims, there is the potential for misascertainment. We suspect that errors would directionally favor underascertainment of events in the subset of individuals with minimal or incomplete representation in clinical-practice data. Indeed, we found that 15% of participants with
known prostate cancer were not represented in clinical-practice data, and nearly 50% of those who underwent RP testing did not have claims-based evidence of RP. Using clinical-practice data aggregated from multiple sources reveals trade-offs in terms of increased sample size but at the potential expense of gaps in data. We sought to address this well-recognized limitation in claims by applying multiple restrictions to select subsets of individuals with evidence of representation in clinical-practice data and by applying hierarchies of selection criteria for identifying the metastasis outcome. Linkages of more than 1 form of clinical data may increase the risks of reidentification. We took numerous steps to ensure that participant privacy was maintained, including the use of anonymized linkages from a third party that enabled no direct transfer of patient identifiers and further deidentification before any research analyses were conducted. Additional limitations include the small proportion of participants with structured EHR data, which limited the sample size for our analyses of changes in PSA values in relation to treatment and outcome events.

Conclusions

We established the first national-scale linkage of transcriptomic and longitudinal clinical data yielding high accuracy for identifying key clinical junctures, including diagnosis, treatment, and early cancer outcome. These findings highlight the potential of integrating transcriptomic and clinical-practice data to enhance understandings of disease biology, patterns of care, and treatment effectiveness.

ARTICLE INFORMATION

Accepted for Publication: April 17, 2024.
Published: June 14, 2024. doi: 10.1001/jamanetworkopen.2024.17274
Open Access: This is an open access article distributed under the terms of the CC-BY-NC-ND License. © 2024 Leapman MS et al. JAMA Network Open.

Corresponding Author: Michael S. Leapman, MD, MHS, Department of Urology, Yale University School of Medicine, 310 Cedar St, BML 238c, New Haven, CT 06520 (michael.leapman@yale.edu).

Author Affiliations: Department of Urology, Yale University School of Medicine, New Haven, Connecticut (Leapman, Martin, Sprengle); Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, Connecticut (Leapman); Veracyte, Inc, San Francisco, California (Ho, Liu, Zhao, Hakansson, Proudfoot, Davicioni); Department of Urology, Emory School of Medicine, Atlanta, Georgia (Filson); Department of Therapeutic Radiology, Yale School of Medicine, New Haven, Connecticut (An); Department of Radiation Medicine and Applied Sciences, University of California, San Diego, La Jolla (Seibert); Department of Radiology, University of California, San Diego, La Jolla (Seibert); Department of Bioengineering, University of California, San Diego, La Jolla (Seibert); Department of Urology, University of Washington, Seattle (Lin); Department of Radiation Oncology, Case Western Reserve University, Cleveland, Ohio (Spratt); Department of Urology, University of California, San Francisco, San Francisco (Cooperberg); Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco (Cooperberg); Department of Urology, Northwestern University Feinberg School of Medicine, Chicago, Illinois (Ross).

Author Contributions: Drs Ho and Liu had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Leapman, Ho, Liu, Davicioni, Lin.
Acquisition, analysis, or interpretation of data: Leapman, Ho, Liu, Filson, Zhao, Hakansson, Proudfoot, Davicioni, Martin, An, Seibert, Spratt, Cooperberg, Ross, Sprengle.
Drafting of the manuscript: Leapman, Liu, Hakansson, Proudfoot.
Critical review of the manuscript for important intellectual content: All authors.
Statistical analysis: Leapman, Ho, Liu, Zhao, Hakansson, Proudfoot.
Obtained funding: Liu, Davicioni.
Administrative, technical, or material support: Leapman, Liu, Davicioni, An, Ross.
Supervision: Leapman, Liu, Davicioni, Spratt, Ross.
Conflict of Interest Disclosures: Dr Liu reported receiving personal fees from Veracyte, Inc, outside the submitted work. Dr Zhao reported receiving Veracyte, Inc, stocks. Dr Halansson reported receiving stock from Veracyte, Inc. Dr Davicioni reported having a patent pending for Veracyte, Inc (US201762448921). Dr Seibert reported receiving personal fees from Varian, WebMD, and GE Healthcare; personal fees and stock options from Cortechs Labs; and grants from GE Healthcare outside the submitted work. Dr Spratt reported receiving personal fees from Astellas, AstraZeneca, Boston Scientific, Janssen, Novartis, and Pfizer outside the submitted work. Dr Cooperberg reported receiving personal fees from Veracyte, MDx Health, Biomarker, Janssen, Merck, Pfizer, Bayer, Astra Zeneca, LymnDx, and Exosome outside the submitted work. Dr Ross reported receiving personal fees from Veracyte for consulting and speaking during the conduct of the study. No other disclosures were reported.

Data Sharing Statement: See Supplement 2.

REFERENCES


SUPPLEMENT 1.
eTable 1. Code list used to create prostate cancer event
eTable 2. Proportion of subjects with prostate cancer events identified by text query versus administrative claims codes
eTable 3. Sources of data from linkage of genomic classifier and real-world clinical data
eTable 4. Incidence of metastasis and prostate cancer specific metastasis among subsets of subjects with known NCCN risk group and CAPRA-S score based on clinical data submitted with Decipher testing. 51,461 of 53,871 Biopsy-tested and 19,608 of 39,105 RP-tested patients
eAppendix 1. Supplemental methods
eAppendix 2. Supplemental results

SUPPLEMENT 2.
Data Sharing Statement