Incorporating Structured and Unstructured Data Sources to Identify Hereditary Cancer Testing among Veterans with Metastatic Castration-Resistant Prostate Cancer

Danielle Candelieri-Surette, MPH[1], Anna Hung, PharmD, PhD, MS[2,3,4], Sally MacDonald, RN (retired)[5], Fatai Agiri, BS[4], Mengke Hu, PhD[1,5], Elizabeth Hanchrow, RN, MSN[1], Kyung Min Lee, PhD[1], Nai-Chung Nelson Chang, PhD[1], Ming Yin, MS[1], Jeffrey W. Shevach, MD[6], Weiyan Li, PhD, MPH[7], Tyler J. Nelson,BS[1], Anthony Gao, MS[1], Kathryn M. Pridgen, MA[1,] Martin W. Schoen, MD, MPH[8,9], Scott L. DuVall, PhD[1,5], Yu-Ning Wong, MD, MSCE[10,11], Julie A. Lynch, PhD, MBA, RN[1,5,*] and Patrick R. Alba, MS[1,5*] (co-senior authors)

[1] VA Informatics and Computing Infrastructure (VINCI), VA Salt Lake City Health Care System, Salt Lake City, UT, USA
[2] Durham VA Medical Center, Durham, NC, USA
[3] Duke Clinical Research Institute, Duke University School of Medicine, Durham, NC, USA
[4] Department of Population Health Sciences, Duke University School of Medicine, Durham, NC, USA
[5] Department of Internal Medicine, Division of Epidemiology, University of Utah School of Medicine, Salt Lake City, UT
[6] Division of Medical Oncology, The Duke Cancer Institute Center for Prostate and Urologic Cancers, Duke University
[7] AstraZeneca Pharmaceuticals, LP, Gaithersburg, MD, USA
[8] Medicine Service, St. Louis Veterans Affairs Health Care System, Saint Louis, MO
[9] Department of Internal Medicine, Division of Hematology/Oncology, Saint Louis University School of Medicine, Saint Louis, MO
[10] Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA, USA
[11] Division of Hematology/Oncology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA

Background

Identifying documentation of hereditary cancer genetic testing is challenging due to poor documentation and complexity in distinguishing germline from somatic testing. This study introduces an integrated approach using structured and unstructured data from the VA national electronic health record to identify and describe patient utilization of hereditary cancer genetic testing among patients with metastatic castration resistant prostate cancer (mCRPC).

Methods

The study population included 9,852 Veterans diagnosed with mCRPC between January 1, 2016, and December 31, 2021, as identified using a previously validated algorithm.

Structured data was gathered from GDx lab data feeds (extracted via PDF downloads), the VA Laboratory file, clinical notes, and the VA structured data domain.

A training dataset for the NLP model was created by manually annotating 245 complete patient charts, comprising 965 distinct notes. We chose notes that were created on dates surrounding billing codes relevant to germline tests or contained any original keywords related to HRR gene names (i.e., *BRCA1, BRCA2, BRCA, ATM, BARD1, CDK12, CHEK1, CHEK2, FANCL, PALB2, RAD51B, RAD51D, RAD54L, BRIP1*) or germline testing. They were comprehensively reviewed to identify genes, tests, laboratories, and their respective statuses. Inter-annotator agreement (IAA) assessment yielded a Cohen's kappa of 0.83. 150 of these annotated charts were used to train the rule-based NLP model.

The NLP model's lexicon was built using terms identified in the training set and supplemented with terms from the NIH Genetic Testing Registry. The algorithm first identifies a lexicon of concepts related to gene names (e.g., "BRCA1," "CHEK2"), genetic test names (e.g., "Invitae Multi-Cancer Gene Panel," "Prostate hereditary cancer panel"), or laboratories (e.g., "Ambry Genetics," "Fulgent"). Contextual rules are then applied to classify the status of the concepts identified (e.g., "positive," "negative," "ordered," "declined," "discussed").

Generalizability and performance were assessed using a charts held out from the development process.

Results

Among the 9,852 Veterans with mCRPC, 1,847 (18.75%) patients had evidence of being offered testing (tested, ordered, or declined); 149 (8.07%) patients declined, and 110 (1.12%) had documentation within the patient notes that the provider would order the test, but lacked evidence that the test order was executed. Among the 1,588 (16.4%) who received genetic testing, 326 were identified from VA clinical notes, 354 from structured data, and 908 from overlapping sources. Figure 1 shows a more detailed breakdown. Only 294 patients had genetic results available.

The performance assessment revealed a positive predictive value of 0.923, sensitivity of 0.857, and F1 score of 0.889.

Conclusions

With this integrated informatics approach, we can differentiate more reliably between hereditary and somatic cancer testing. This advancement will facilitate longitudinal studies to assess the long-term outcomes of patients who undergo genetic testing, particularly those identified as high-risk or with aggressive diseases.

Funding Acknowledgements

This study was funded by AstraZeneca and Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA, who are co-developing olaparib.

Conflicts of Interest Disclosure Statement

AH reports support from Career Development Award Number IK2 HX003359 from the United States Department of Veterans Affairs Health Services Research & Development Service. WL is an employee of and owns stock for AstraZeneca Pharmaceuticals, LP (Gaithersburg, MD, USA). JAL, SLD, and PRA report grants from Alnylam Pharmaceuticals, Inc., Astellas Pharma, Inc., AstraZeneca Pharmaceuticals LP, Biodesix, Inc, Celgene Corporation, Cerner Enviza, GSK PLC, IQVIA Inc., Janssen Pharmaceuticals, Inc., Novartis International AG, Parexel International Corporation through the University of Utah or Western Institute for Veteran Research outside the submitted work.

Acknowledgements

This material is the result of work supported with resources and the use of facilities at the VA Salt Lake City Healthcare System and Durham VA Medical Center. The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government.

Figure 1. Patients who received germline testing who were identified from each data source