Association of Pathogenic Variants in Hereditary Cancer Genes With Multiple Diseases

Chenjie Zeng, PhD, MPH; Lisa A. Bastarache, MS; Ran Tao, PhD; Eric Venner, PhD; Scott Hebbringer, PhD; Justin D. Andujar, BS; Sarah T. Bland, MPH, MBA; David R. Crosslin, PhD; Siddharth Pratap, PhD, MS; Ayorinde Cooley, BS; Jennifer A. Pacheco, MS; Kurt D. Christensen, PhD; Emma Perez, MS; Carrie L. Blout Zawatsky, MS; Leora Witkowski, PhD; Hana Zouk, PhD; Chunhua Weng, PhD; Kathleen A. Leppig, MD; Patrick M. A. Sleiman, PhD; Hakon Hakonarson, MD, PhD; Marc S. Williams, MD; Yuan Luo, PhD; Gail P. Jarvik, MD, PhD; Patrick M. A. Sleiman, PhD; Hakon Hakonarson, MD, PhD; Marc S. Williams, MD; Yuan Luo, PhD; Gail P. Jarvik, MD, PhD; Heidemarie Launikonis, MD; Jillian A. Lichtenstein, MD; Scott A. Rogers, MD; David R. Crosslin, PhD; Siddharth Pratap, PhD, MS; Ayorinde Cooley, BS; Jennifer A. Pacheco, MS; Kurt D. Christensen, PhD; Emma Perez, MS; Carrie L. Blout Zawatsky, MS; Leora Witkowski, PhD; Hana Zouk, PhD; Chunhua Weng, PhD; Kathleen A. Leppig, MD; Patrick M. A. Sleiman, PhD; Hakon Hakonarson, MD, PhD; Marc S. Williams, MD; Yuan Luo, PhD; Gail P. Jarvik, MD, PhD; Robert C. Green, MD, MPH; Wendy K. Chung, MD; Ali G. Gharavi, MD; Niall J. Lennon, PhD; Heidi L. Rehm, MD; Richard A. Gibbs, PhD; Josh F. Peterson, MD, MPH; Dan M. Roden, MD; Georgia L. Wiesner, MD, MS; Joshua C. Denny, MD, MS

IMPORTANCE Knowledge about the spectrum of diseases associated with hereditary cancer syndromes may improve disease diagnosis and management for patients and help to identify high-risk individuals.

OBJECTIVE To identify phenotypes associated with hereditary cancer genes through a phenome-wide association study.

DESIGN, SETTING, AND PARTICIPANTS This phenome-wide association study used health data from participants in 3 cohorts. The Electronic Medical Records and Genomics Sequencing (eMERGEseq) data set recruited predominantly healthy individuals from 10 US medical centers from July 16, 2016, through February 18, 2018, with a mean follow-up through electronic health records (EHRs) of 12.7 (7.4) years. The UK Biobank (UKB) cohort recruited participants from March 15, 2006, through August 1, 2010, with a mean (SD) follow-up of 12.4 (1.0) years. The Hereditary Cancer Registry (HCR) recruited patients undergoing clinical genetic testing at Vanderbilt University Medical Center from May 1, 2012, through December 31, 2019, with a mean (SD) follow-up through EHRs of 8.8 (6.5) years.

EXPOSURES Germline variants in 23 hereditary cancer genes. Pathogenic and likely pathogenic variants for each gene were aggregated for association analyses.

MAIN OUTCOMES AND MEASURES Phenotypes in the eMERGEseq and HCR cohorts were derived from the linked EHRs. Phenotypes in UKB were from multiple sources of health-related data.

RESULTS A total of 214,020 participants were identified, including 23,544 in eMERGEseq cohort (mean [SD] age, 47.8 [23.7] years, 12,611 women [53.6%]), 187,234 in the UKB cohort (mean [SD] age, 56.7 [8.1] years, 104,055 [55.6%] women), and 3,242 in the HCR cohort (mean [SD] age, 52.5 [15.5] years, 2,851 [87.9%] women). All 38 established gene-cancer associations were replicated, and 19 new associations were identified. These included the following 7 associations with neoplasms: CHEK2 with leukemia (odds ratio [OR], 3.81 [95% CI, 2.64-5.48]) and plasma cell neoplasms (OR, 3.12 [95% CI, 1.84-5.28]), ATM with gastric cancer (OR, 4.27 [95% CI, 2.35-7.44]) and pancreatic cancer (OR, 4.44 [95% CI, 2.66-7.40]), MUTYH (biallelic) with kidney cancer (OR, 32.28 [95% CI, 6.40-162.73]), MSH6 with bladder cancer (OR, 5.63 [95% CI, 2.75-11.49]), and APC with benign liver/intrahepatic bile duct tumors (OR, 5.21 [95% CI, 1.49-18.19]). The remaining 12 associations with nonneoplastic diseases included BRCA1/2 with ovarian cysts (OR, 3.15 [95% CI, 2.22-4.46] and 3.12 [95% CI, 2.36-4.12], respectively), MEN1 with acute pancreatitis (OR, 33.45 [95% CI, 9.25-121.02]), APC with gastritis and duodenitis (OR, 4.66 [95% CI, 2.61-8.33]), and PTEN with chronic gastritis (OR, 15.68 [95% CI, 6.01-40.92]).

CONCLUSIONS AND RELEVANCE The findings of this genetic association study analyzing the EHRs of 3 large cohorts suggest that these new phenotypes associated with hereditary cancer genes may facilitate early detection and better management of cancers. This study highlights the potential benefits of using EHR data in genomic medicine.

JAMA Oncol. doi:10.1001/jamaoncol.2022.0373
Published online April 21, 2022.

© 2022 American Medical Association. All rights reserved.
Understanding the phenotypic consequences of genomic variation is critical to genomic medicine. Uncovering gene-phenotype associations facilitates clinical diagnoses, leads to better treatment, improves prognosis, and provides insights into disease etiology and potential therapeutic targets. The application of next-generation sequencing has markedly accelerated the discovery of novel mendelian disease genes and has expanded our knowledge of their characteristic phenotypes. These are epitomized by hereditary cancer genes. Their associated phenotypes have been shown to extend beyond predisposition to cancer. However, substantial gaps in knowledge about the spectrum of phenotypes have been noted, suggesting the need for infrastructure and resources to systematically assess gene-phenotype associations.

Current approaches to uncover phenotypes include family-based and population-based studies, most of which focused on 1 gene and/or 1 trait or similar traits. These studies have fundamentally improved our understanding of diseases and laid foundations for precision medicine. Systematic efforts to collect information on gene-phenotype associations include the Online Mendelian Inheritance in Man (OMIM), which curates knowledge through literature review with decades of efforts.

In this study, we hypothesize that additional conditions are associated with hereditary cancer genes. Using an alternative approach, namely, the phenome-wide association study (PhE WAS), we used the phenotypic data derived from health record data from 3 cohorts, totaling 214,020 participants, to investigate a broad range of phenotypes associated with hereditary cancer genes.

**Methods**

Two clinical site–based cohorts and 1 population-based prospective cohort were included in this PhE WAS. All US-based studies were approved by local institutional review boards, and the UK-based study was approved by relevant ethics committees and organizations. Details are provided in eMethods in the Supplement. All participants provided written informed consent according to approved protocols. This study followed the Strengthening the Reporting of Genetic Association Studies (STREGA) guideline.

**Study Populations**

The Electronic Medical Records and Genomics Sequencing (eMERGEseq) cohort consisted of 24,956 biobank and prospectively recruited predominantly healthy individuals from 10 clinical sites under the eMERGE network from July 16, 2016, through February 18, 2018. The primary goal of this project was to provide clinical genetic testing and return actionable genetic results to patients. A total of 52% of the participants were unselected and mainly recruited from primary care clinics or identified from biobanks without specific indications, with the others recruited from specific clinics depending on site-specific interests. A detailed description of each site, including enrollment criteria, specific research interest, and enrichment of phenotypes is provided in eMethods in the Supplement.

For this study, we removed individuals without International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), or International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Clinical Modification (ICD-10-CM), codes in the electronic health records (EHRs). A total of 23,544 individuals were retained for analysis.

The Hereditary Cancer Registry (HCR) at Vanderbilt University Medical Center included all 3,794 individuals who received clinical genetic testing for hereditary cancer from May 1, 2012, through December 31, 2019, and who agreed to and consented to be included in this registry. Results of genetic testing were documented in the EHRs. We obtained the EHR data of 3,739 individuals through the Research Derivative, a database of clinical and related data derived from EHR systems. Through reviewing clinical records in the HCR, we removed patients who were also participants of the eMERGEseq project and family members of the index patients who were enrolled in the registry owing to cascade testing (n = 483). A total of 3,242 patients were retained for analyses.

The UK Biobank (UKB) is a prospective population-based cohort of 300,217 participants recruited from March 15, 2006, through August 1, 2010, who are continuously followed up. We included 200,619 participants with whole-exome sequencing data available at the time of this study. After removing related participants (n = 5,007) and those without ICD-9 and ICD-10 codes (n = 8,378), 187,234 participants were included.

**Sequencing and Variant Classification**

Germline variant data in the eMERGEseq cohort were obtained from targeted sequencing. Details on the design of the sequencing panel have been described previously. Briefly, this panel consists of a total of 109 genes, including 58 genes from the American College of Medical Genetics and Genomics (ACMG) actionable finding list and 51 genes nominated by participating sites. The full list of these genes is provided in eTable 1 in the Supplement. Among the 58 genes from the ACMG panel, we selected all genes (n = 25) determined to be associated with cancer phenotypes by the ACMG Secondary Findings Working Group. These genes were APC (OMIM 61731), BMPRIA (OMIM 601299), BRCA1 (OMIM 113705),
Association of Pathogenic Variants in Hereditary Cancer Genes With Multiple Diseases

Original Investigation Research

BRCA2 (OMIM 600185), MEN1 (OMIM 613733), MLHI (OMIM 120436), MSH2 (OMIM 609309), MSH6 (OMIM 600678), MUTYH (OMIM 604933), NF2 (OMIM 607379), PMS2 (OMIM 600259), PTEN (OMIM 601728), RB1 (OMIM 164041), RET (OMIM 164761), SDHA (OMIM 613019), SDHB (OMIM 185470), SDHC (OMIM 602413), SDHD (OMIM 602690), SMAD4 (OMIM 600993), STK11 (OMIM 602690), TP53 (OMIM 191170), TSC1 (OMIM 605284), TSC2 (OMIM 191092), VHL (OMIM 608537), and WT1 (OMIM 607102). We also included genes related to cancer phenotypes determined by field experts in the eMERGE network from genes selected by participant sites. These genes were ATM (OMIM 607585), BLM (OMIM 604610), CHEK2 (OMIM 604373), PALB2 (OMIM 610355), POLE (OMIM 174761), and POLE (OMIM 174762). Genetic testing for patients in the HCR was performed by commercial Clinical Laboratory Improvement Amendments (CLIA)- and College of American Pathologists (CAP)-accredited molecular diagnostic laboratories. Results from CLIA- and CAP-accredited laboratories were considered highly accurate.26 Germline variant data in UKB were obtained by whole-exome sequencing data through the UKB data center as described elsewhere.27

Variant classification in eMERGEseq was performed by 2 CLIA- and CAP-accredited laboratories according to ACMG and Association for Medical Pathology guidelines with modifications by experts as previously described.19 Variant classification in the HCR was performed by commercial CLIA- and CAP-accredited molecular genetic testing laboratories. Variant classification in the UKB was performed according to the ACMG and ClinGen guidelines.19 With the exception of APC11307K, which was classified as a risk allele, all detected variants were classified into pathogenic, likely pathogenic, variant of uncertain significance, likely benign, and benign. We compared results of shared variants and made the classifications identical across all studies according to the aforementioned guidelines. Details are described in eMethods in the Supplement.

For each gene, we defined patients with pathogenic and likely pathogenic variants as carriers and patients with no rare variants or only benign or likely benign variants as noncarriers, and patients with variants of uncertain significance as carriers of these variants. Only genes with at least 10 carriers in all cohorts combined were included. For MUTYH, only biallelic variant carriers were considered. The frequency of carriers in each cohort was consistent with that in previous studies with similar settings.28–36

PheWAS Phenotypes
In the eMERGEseq and HCR cohorts, we extracted the ICD-9-CM and ICD-10-CM data from linked EHRs. The validity of this EHR-based PheWAS approach has been demonstrated in previous studies.35,37–43 In the UKB, we extracted ICD-9 and ICD-10 data from the harmonized health outcome data derived from cancer and death registries, inpatient medical records, and self-reported health outcomes.23 Details on the clinical data linkage and standardized questionnaires and interviews have been described previously.23 Specifically, self-reported cancer diagnoses were validated against data from cancer registries and mapped to ICD-10 codes. Noncancer self-reported health outcomes were also mapped to ICD-10 codes if applicable. Dates of first occurrence of diseases were also extracted. We mapped all ICD codes to phecodes to define the phenotypes for the PheWAS.37,18,44 Details are provided in eMethods in the Supplement. A total of 3483 unique phecodes were derived from the eMERGEseq data set; 2853, from the HCR data set; and 2693, from the UKB data set. These phecodes covered 15 categories of diseases and conditions, including congenital, cardiovascular, dermatologic, developmental, digestive, endocrine, hematopoietic, infectious, neoplastic, pregnant, psychiatric, pulmonary, genitourinary, muscular-skeletal, and symptoms and/or signs.

Statistical Analysis
Statistical analysis was conducted from April 2020 to October 2021. We performed gene-level association tests by collapsing pathogenic and likely pathogenic variants in the same gene. We removed participants with variants of uncertain significance in the same gene from analyses. Each gene-phenotype association was tested independently using the Firth logistic regression.45,46 In the eMERGEseq cohort, we adjusted for age, EHR length in years, sites, the first 4 principal components, and sex if applicable. In the HCR cohort, we adjusted for age, EHR length in years, self-reported race, and sex if applicable. In the UKB cohort, we adjusted for age, length of follow-up in years, sites, the first 16 principal components suggested by Privé et al,47 and sex if applicable. The race variable (defined by principal components or self-reported) was included to account for population stratifications. Analyzes were performed assuming an autosomal dominant inheritance for all genes except MUTYH, for which an autosomal recessive inheritance was assumed, according to their inheritance patterns as hereditary cancer genes documented in the OMIM database,15 the comprehensive, authoritative collection of gene-phenotype correlations. The association of monoallelic MUTYH variants with cancer remains inconclusive;48 therefore, we did not perform analyses for these carriers. All populations were included. The number of phenotypes evaluated in each cohort is presented in eFigure 1 in the Supplement. We only considered associations found in at least 2 of the 3 studies with the same direction of effect. Meta-analyses were performed assuming a fixed-effect model. We defined 2.5 × 10−5 as the empirical phenome-wide significance threshold at a significance level of α = .05 through permutations (eFigure 2 and eMethods in the Supplement). All statistical analyses were performed using R, version 4.0.1 (R Project for Statistical Computing). Figure 2 was produced by the R package ggsanatogram,49 which uses the tissue coordinates from the Expression Atlas.50

We categorized all gene associations into 3 groups: known or primary associations as documented in the OMIM database, associations related to known phenotypes (eg, elevated cancer antigen 125 for BRCA1/2), and potentially new associations. We considered that a known phenotype-gene association was replicated in our analysis if the PheWAS had a P < .05 with the expected direction of the effect. Details are described in eMethods in the Supplement.

Sensitivity Analysis
We conducted several sensitivity analyses to test the robustness of the new associations. First, we tested associations lim-
iting the individuals to those with European ancestry. For the eMERGEseq and UKB cohorts, we also derived ancestral specific principal components from genetic data as additional covariates. Second, to investigate whether associations with noncancer phenotypes were associated with prior cancer diagnoses, we restricted analyses to those without any cancer diagnoses (excluding basal cell carcinomas) before enrollment. We used the UKB data set because dates of cancer diagnoses were ascertained through cancer registries. Second, to investigate whether the observed associations between CHEK2 and hematological malignant neoplasms were associated with prior cancer diagnoses, we restricted analyses to those without cancer diagnoses before blood sampling in the UKB data set. In addition, we conducted another analysis by removing participants with any cancer diagnosis within 3 years after blood sampling. Third, to investigate whether the observed association between CHEK2 and leukemia differed by subtype, we evaluated associations of CHEK2 with subtypes of leukemia. Finally, we compared association results of BRCA1 with BRCA2 found in this PheWAS.

### EHR Reviews

We conducted EHR reviews for participants with readily accessible EHRs at Vanderbilt University Medical Center to gather more information about diagnoses related to new associations. We verified diagnoses by reviewing pathology reports, radiology imaging results, and clinical narratives.

### Results

We included 214,020 participants from the 3 cohorts. Baseline demographic characteristics and the follow-up time for each cohort are summarized in eTable 2 in the Supplement. Participants in the eMERGEseq cohort (n = 23,544) had a mean (SD) age of 47.8 (23.7) years with a mean (SD) follow-up of 12.7 (7.4) years; 12,611 (53.6%) were women, and 10,933 (46.4%) were men. The HCR cohort (n = 3,242) had a mean (SD) age of 52.5 (15.5) years with a mean (SD) follow-up of 8.8 (6.5) years; 413 (12.7%) had non-European ancestry, 2,851 (87.9%) were women, and 391 (12.1%) were men. The UKB cohort (n = 187,234) had a mean (SD) age at recruitment of 56.7 (8.1) years with a mean (SD) follow-up time of 12.4 (1.0) years; 11,293 (6.0%) had non-European ancestry; 104,055 (55.6%) were women and 83,179 (44.4%) were men. The distribution of carriers for each gene is summarized in Table 1. A total of 858 carriers were identified in the eMERGEseq cohort; 434, in the HCR cohort; and 5,223 in the UKB cohort. The distribution of carriers for each gene by ancestral group is summarized in eTable 3 in the Supplement.

We first assessed whether the PheWAS could replicate known gene-phenotype associations. Our PheWAS replicated 38 of 38 primary gene-cancer associations (100%) and 164 of 235 gene-phenotype associations (69.8%) documented in OMIM, which reports diseases and symptoms as-

---

**Table 1. Number of Carriers and Noncarriers in the eMERGEseq, HCR, and UKB Cohorts**

<table>
<thead>
<tr>
<th>Gene</th>
<th>eMERGEseq cohort (n = 23,544)</th>
<th>HCR cohort (n = 3,242)</th>
<th>UKB cohort (n = 187,234)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of carriers</td>
<td>No. of noncarriers</td>
<td>Carriers, %</td>
<td>No. of carriers</td>
</tr>
<tr>
<td>APC</td>
<td>14</td>
<td>21,903</td>
<td>0.06</td>
</tr>
<tr>
<td>ATM</td>
<td>82</td>
<td>21,663</td>
<td>0.34</td>
</tr>
<tr>
<td>BRCA1</td>
<td>82</td>
<td>23,012</td>
<td>0.34</td>
</tr>
<tr>
<td>BRCA2</td>
<td>138</td>
<td>22,325</td>
<td>0.58</td>
</tr>
<tr>
<td>CHEK2</td>
<td>272</td>
<td>22,697</td>
<td>1.10</td>
</tr>
<tr>
<td>MEN1</td>
<td>2</td>
<td>23,269</td>
<td>0.01</td>
</tr>
<tr>
<td>MLH1</td>
<td>14</td>
<td>23,214</td>
<td>0.06</td>
</tr>
<tr>
<td>MSH2</td>
<td>16</td>
<td>22,361</td>
<td>0.06</td>
</tr>
<tr>
<td>MSH6</td>
<td>50</td>
<td>22,634</td>
<td>0.21</td>
</tr>
<tr>
<td>MUTYH (biallelic)</td>
<td>4</td>
<td>22,436</td>
<td>0.02</td>
</tr>
<tr>
<td>PALB2</td>
<td>28</td>
<td>22,925</td>
<td>0.13</td>
</tr>
<tr>
<td>PMS2</td>
<td>54</td>
<td>22,300</td>
<td>0.23</td>
</tr>
<tr>
<td>PTEN</td>
<td>13</td>
<td>23,184</td>
<td>0.06</td>
</tr>
<tr>
<td>RB1</td>
<td>2</td>
<td>23,105</td>
<td>0.01</td>
</tr>
<tr>
<td>RET</td>
<td>34</td>
<td>22,745</td>
<td>0.14</td>
</tr>
<tr>
<td>SDHC</td>
<td>6</td>
<td>23,397</td>
<td>0.02</td>
</tr>
<tr>
<td>SDHD</td>
<td>6</td>
<td>23,340</td>
<td>0.02</td>
</tr>
<tr>
<td>SDHC</td>
<td>4</td>
<td>23,435</td>
<td>0.02</td>
</tr>
<tr>
<td>TPS3</td>
<td>12</td>
<td>23,336</td>
<td>0.06</td>
</tr>
<tr>
<td>TSC1</td>
<td>5</td>
<td>22,792</td>
<td>0.02</td>
</tr>
<tr>
<td>TSC2</td>
<td>12</td>
<td>21,598</td>
<td>0.05</td>
</tr>
<tr>
<td>VHL</td>
<td>5</td>
<td>23,329</td>
<td>0.02</td>
</tr>
<tr>
<td>WT1</td>
<td>3</td>
<td>23,215</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Abbreviations: eMERGEseq, Electronic Medical Records and Genomics Sequencing; HCR, hereditary cancer registry; UKB, UK Biobank.
The probability of replicating associations in at least 164 of 235 tests by chance, under the null hypothesis of no association, is $P = 2.13 \times 10^{-154}$. A total of 193 gene-phenotype associations exceeded the phenome-wide significance ($P < 2.5 \times 10^{-5}$). After removing known associations, 19 new associations that have not been documented in the OMIM database were found in 13 hereditary cancer genes (Table 2 and Figure 2). These consisted of 6 associations with malignant tumors, including CHEK2 with leukemia (odds ratio [OR], 3.81 [95% CI, 2.64-5.48]) and plasma cell neoplasms (OR, 3.12 [95% CI, 1.84-5.28]), ATM with gastric cancer (OR, 4.27 [95% CI, 2.35-7.44]) and pancreatic cancer (OR, 4.44 [95% CI, 2.66-7.40]), MUTYH (biallelic) with kidney cancer (OR, 32.28 [95% CI, 6.40-162.73]), MSH6 with bladder cancer (OR, 5.63 [95% CI, 2.75-11.49]), and an association of APC with benign liver/intrahepatic bile duct tumors (OR, 52.01 [95% CI, 14.29-189.29]). Ten genes were associated with nonneoplastic diseases (eg, BRCA1 [OR, 3.15 [95% CI, 2.22-4.46]] and BRCA2 [OR, 3.12 [95% CI, 2.36-4.12]] with ovarian cysts, MEN1 with acute pancreatitis [OR, 33.45 [95% CI, 9.25-121.02]], APC with gastritis and duodenitis [OR, 4.66 [95% CI, 2.61-8.33]], and PTEN with chronic gastritis [OR, 15.68 [95% CI, 6.01-40.92]]).

All results of the sensitivity analyses were consistent with the main findings. Results of new associations remained largely unchanged in the analyses conducted in European descendants only (eTable 5 in the Supplement). After removing participants with prior cancer diagnoses, associations between BRCA1/2 and ovarian cyst, PTEN and chronic gastritis, and MEN1 with acute pancreatitis remained statistically significant (eTable 6 in the Supplement). For associations of CHEK2 with hematological cancers, removing participants with prior cancer diagnoses or even those with cancer diagnoses within 3 years after blood draw did not substantially change the associations (eTable 7 in the Supplement). No substantial differences were detected in the associations of CHEK2 with subtypes of leukemia (eTable 8 in the Supplement).
By reviewing EHRs of participants in the HCR cohort, we verified the diagnosis of renal cell carcinoma in the biallelic MUTYH variant carrier, which was consistent with the diagnosis of this cancer in the biallelic MUTYH variant carrier in the UKB cohort. We found a diagnosis of thyroid cancer in the RET carrier with diplopia. We also found that 7 of 20 BRCA1 carriers with ovarian cysts were diagnosed with ovarian cancer, whereas only 2 such diagnoses were found in BRCA2 carriers with ovarian cysts (2 of 24). However, the difference between BRCA1 and BRCA2 was not statistically significant (P = .06). We also did not find evidence that patients who were BRCA1/2 carriers with ovarian cysts were actually cases of ovarian cancer that had been missed. We did not find pancreatic cancer diagnoses among MEN1 carriers with acute pancreatitis.

Discussion

In this study, we demonstrate the feasibility of rapid phenotype discovery by the PheWAS approach by replicating most known associations documented in the OMIM that represented knowledge accumulated in decades. We identified 19 new gene-phenotype associations, which spanned categories of diseases, including the neoplastic, genitourinary, digestive, congenital, metabolic, psychiatric, and neurological categories, supporting our hypothesis that hereditary cancer syndromes are associated with multiple diseases.

This study revealed a novel gene-cancer association between biallelic MUTYH variants and kidney cancer. Although monogenic germline MUTYH pathogenic variants have been identified in patients with renal cell carcinoma, previous studies that used data from high-risk families and probands reported no occurrence of this cancer among biallelic MUTYH variant carriers but did report benign kidney lesions. Although few studies have investigated the role of MUTYH in kidney cancers, some mutation signatures of genomic instability have been found to be more common in these tumors than other solid tumors. Further molecular studies are needed to illuminate the observed association.

This study provides additional evidence for associations of cancers that have not been documented in the OMIM data.

Table 2. New Associations Discovered via PheWAS

<table>
<thead>
<tr>
<th>Gene</th>
<th>Phenotype</th>
<th>Cohort, OR (95% CI)</th>
<th>HCR</th>
<th>UKB</th>
<th>Meta-analysis</th>
<th>P value Summary</th>
<th>Het</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATM</td>
<td>Pancreatic cancer</td>
<td>2.54 (0.39-16.39)</td>
<td>3.68 (0.77-17.52)</td>
<td>4.79 (2.72-8.43)</td>
<td>4.44 (2.66-7.40)</td>
<td>7.88 × 10−8</td>
<td>.79</td>
</tr>
<tr>
<td>ATM</td>
<td>Gastric cancer</td>
<td>4.62 (0.61-34.79)</td>
<td>NA</td>
<td>4.24 (2.27-7.90)</td>
<td>4.27 (2.35-7.74)</td>
<td>1.80 × 10−6</td>
<td>.94</td>
</tr>
<tr>
<td>CHEK2</td>
<td>Leukemia</td>
<td>4.42 (2.18-8.94)</td>
<td>5.04 (1.00-25.41)</td>
<td>3.52 (2.26-5.47)</td>
<td>3.81 (2.64-5.48)</td>
<td>6.18 × 10−12</td>
<td>.81</td>
</tr>
<tr>
<td>CHEK2</td>
<td>Plasma cell neoplasmsa</td>
<td>2.66 (0.90-7.90)</td>
<td>NA</td>
<td>3.28 (1.79-5.98)</td>
<td>3.12 (1.84-5.28)</td>
<td>2.30 × 10−5</td>
<td>.74</td>
</tr>
<tr>
<td>MSH6</td>
<td>Bladder cancer</td>
<td>8.30 (3.29-25.94)</td>
<td>18.98 (4.32-83.30)</td>
<td>2.28 (0.79-6.61)</td>
<td>5.63 (2.75-11.49)</td>
<td>1.33 × 10−5</td>
<td>.06</td>
</tr>
<tr>
<td>MUTYH</td>
<td>Kidney cancer</td>
<td>NA</td>
<td>84.13 (8.47-836.11)</td>
<td>12.57 (1.29-122.74)</td>
<td>32.28 (6.40-162.73)</td>
<td>2.50 × 10−5</td>
<td>.25</td>
</tr>
<tr>
<td>APC</td>
<td>Benign liver/IHBD tumor</td>
<td>61.01 (7.66-485.98)</td>
<td>26.47 (3.48-201.34)</td>
<td>146.80 (16.10-186.83)</td>
<td>52.01 (14.29-189.29)</td>
<td>1.57 × 10−8</td>
<td>.62</td>
</tr>
</tbody>
</table>

Abbreviations: eMERGEseq, Electronic Medical Records and Genomics Sequencing; GI, gastrointestinal tract; HCR, Hereditary Cancer Registry; IHBD, intrahepatic bile duct; NA, not applicable; OR, odds ratio; PheWAS, phenotype-wide association study; UKB, UK Biobank.

a We used Firth logistic regression in this PheWAS assuming a dominant model.

b Plasma cell neoplasms also include multiple myeloma.
base but were reported in previous literature, including \textit{ATM} with gastric and pancreatic cancer,\cite{53,54} \textit{MSH6} with bladder cancer,\cite{55} and \textit{CHEK2} with leukemia.\cite{56-58} Notably, a recent study\cite{59} suggested that loss of \textit{CHEK2} function increased the risk of clonal hematopoiesis of indeterminate potential, which was a risk factor for hematological malignant neoplasms.\cite{60} Furthermore, it was found that prior cancer therapies could increase the risk of clonal hematopoiesis of indeterminate potential.\cite{61} We observed that associations with leukemia or multiple myeloma persisted after removing participants with cancer diagnosed before and within 3 years of blood sampling. Nonetheless, we could not fully exclude the possibility that somatic variants of \textit{CHEK2} detected owing to clonal expansions contribute to the observed association. Future studies that include additional types of tissues can help exclude the somatic variants and validate the observed associations.

This study also revealed new noncancer associations that would have been difficult to detect in studies focusing on cancers or using prior knowledge.\cite{62} These phenotypes included inflammation-related disorders, which were consistent with previous findings.\cite{53,64} For example, a recent study\cite{63} identified an essential role of \textit{MEN1} in exocrine pancreas homeostasis in response to inflammation that contributes to pancreatic tumors in mouse models. A previous study\cite{65} suggested that \textit{MUTYH} contributed to inflammatory-related disorders. We found that homozygous or compound heterozygous \textit{MUTYH} carriers had an association with polycystic ovaries, for which chronic inflammation has been proposed to be a key contributor.\cite{66} We also found a Bechetsyndrome diagnosis in a \textit{MUTYH} biallelic variant carrier in this study. Taken together, these findings provide supporting evidence for a role of \textit{MUTYH} in inflammatory-related disorders.

Results of EHR review suggested that some of the noncancer phenotypes could be symptoms of underlying diseases that had been known. For example, the association of \textit{RET} with diplopia was likely to be mediated by neuroendocrine disorders, including tumors. However, diplopia has been largely underreported in patients with multiple endocrine neoplasia type 2 in previous studies and thus has not been documented in the OMIM database. We believe that recognizing such relevant symptoms can be important for the management of multiple endocrine neoplasia type 2. Identification of symptoms such as these may also serve an early sign of underlying diseases such as cancers and thus facilitate early detection, as shown in previous studies.\cite{67,68}
Owing to the small number of carriers identified for genes including MUTYH, VHL, and APC, additional studies are needed to validate the new associations identified in this study. A previous study\(^6\) suggested that BRCA1/2 could have a different role in diseases. However, we did not observe different associations in our PheWAS. Follow-up studies are needed to test this hypothesis.

**Limitations**

Limitations of this study include a relatively small sample size of populations of non-European descendants. This could potentially limit the generalizability of our findings to these populations, although we included them in analyses. We anticipate that large EHR-based cohorts including more diverse populations, such as the All of Us research program,\(^2\) will identify additional phenotypes associated with these genes and increase the generalizability of the findings to these underrepresented populations.

**ARTICLE INFORMATION**

Accepted for Publication: January 7, 2022.

Published Online: April 21, 2022. doi:10.1001/jamaoncol.2022.0373

**Author Affiliations:** National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland (Zeng, Denny); Center for Precision Medicine, Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee (Bastarache, Bland, Peterson, Roden); Department of Biostatistics, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee (Tao); Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas (Venner, Gibbs); Center for Human Genetics, Marshfield Clinic Research Institute, Marshfield, Wisconsin (Hebbring); Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee (Andujar, Wiesner); Clinical and Translational Hereditary Cancer Program, Division of Genetic Medicine, Vanderbilt-Ingram Cancer Center, Vanderbilt University, Nashville, Tennessee (Andujar, Wiesner); Department of Biomedical Informatics and Medical Education, University of Washington School of Medicine, Seattle (Crosslin); School of Graduate Studies and Research, Meharry Medical College, Nashville, Tennessee (Pratap); Department of Microbiology, Immunology and Physiology, Meharry Medical College, Nashville, Tennessee (Cooley); Center for Genetic Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois (Pacheco); PRECISOn Medicine Translational Research (PROMoTeR) Center, Department of Population Medicine, Harvard Pilgrim Health Care Institute, Boston, Massachusetts (Christensen); Department of Population Medicine, Harvard Medical School, Boston, Massachusetts (Christensen); Division of Genetics, Department of Medicine, Brigham and Women’s Hospital, Boston, Massachusetts (Perez, Zawatsky); Centre Universitaire de Santé McGill, McGill University Health Centre, Montreal, Quebec, Canada (Witkowski); Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, Cambridge, Massachusetts (Zouk); Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston (Zouk);

Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, New York (Wiesner); Services and Kaiser Permanente Washington Health Research Institute, Kaiser Permanente of Washington, Seattle (Leppig); Center for Applied Genomics, Children’s Hospital of Philadelphia, Philadelphia, Pennsylvania (Sleiman, Hakonarson); Division of Human Genetics, Department of Pediatrics, The University of Pennsylvania Perelman School of Medicine, Philadelphia (Sleiman, Hakonarson); Genomic Medicine Institute, Geisinger, Danville, Pennsylvania (Williams); Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois (Luo); Department of Medicine (Medical Genetics), University of Washington, Seattle (Jarvik); Department of Genomics Sciences, University of Washington, Seattle (Jarvik); Brigham and Women’s Hospital, Broad Institute, Ariadne Labs and Harvard Medical School, Boston, Massachusetts (Green); Department of Pediatrics, Columbia University, New York, New York (Chung); Department of Medicine, Columbia University, New York, New York (Chung); Division of Nephrology, Department of Medicine, Columbia University Irving Medical Center, New York, New York (Garavì); Center for Precision Medicine and Genomics, Department of Medicine, Columbia University Irving Medical Center, New York, New York (Garavì); Broad Institute of MIT and Harvard, Cambridge, Massachusetts (Lennon); Medical & Population Genetics Program and Genomics Platform, Broad Institute of MIT and Harvard Cambridge, Cambridge, Massachusetts (Rehm); Center for Genomic Medicine, Massachusetts General Hospital, Boston (Rehm); Department of Pathology, Harvard Medical School, Boston, Massachusetts (Rehm); Divisions of Cardiovascular Medicine and Clinical Pharmacology, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee (Rodan); Department of Pharmacology, Vanderbilt University, Nashville, Tennessee (Rodan).

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

**Conclusions**

In this PheWAS of 3 cohorts using data derived from the EHRs of 214 020 participants, we studied a wide range of phenotypes associated with hereditary cancer genes. We identified 19 new gene-phenotype associations, including both neoplastic and nonneoplastic diseases. These findings suggest that PheWAS in EHR data sets has the potential to expand our knowledge of the phenotypes and disease processes in patients with pathogenic and likely pathogenic variants in hereditary cancer genes. New clinical management protocols could be developed based on these findings, so future research replicating these new associations will be important. Large EHR-based cohorts of diverse populations will help reveal the true clinical spectrum of genetic diseases, aid in variant interpretation, and ultimately facilitate precision medicine for all patients.

**Conflict of Interest Disclosures:** Ms Bastarache reported receiving royalties from Nashville Biotech and consulting for Galatea Bio Inc outside the submitted work. Dr Hebbring reported receiving grants from the National Institutes of Health (NIH) during the conduct of the study. Ms Bland reported receiving grants from the National Human Genomic Research Institute of the NIH (NHGRI) during the conduct of the study. Dr Crosslin reported receiving grants from the NHGRI during the conduct of the study and personal fees from UnitedHealth Group outside the submitted work. Dr Christensen reported receiving grants from Sanford Health and the NIH outside the submitted work. Dr Zouk reported receiving grants from the NHGRI during the conduct of the study. Dr Williams reported receiving grants from NIH during the conduct of the study. Dr Luo reported receiving grants from the NIH during the conduct of the study. Dr Jarvik reported receiving grants from NIH during the conduct of the study.
Association of Pathogenic Variants in Hereditary Cancer Genes With Multiple Diseases

Role of the Funders/Sponsors: This study was supported by grant R01LM010685 from the National Library of Medicine and funding to the eMERGE sites through the following: grants U01HG6857, U01HG006375, and U01HG004610 (Kaiser Permanente Washington/University of Washington); U01HG6867 (Bingham and Women’s Hospital); U01HG6872, U01HG006378, and U01HG004608 (Vanderbilt University Medical Center); U01HG6666 and U01HG006628 (Cincinnati Children’s Hospital Medical Center); U01HG6379 and U01HG04599 (Mayo Clinic); U01HG6879 and U01HG006382 (Geisinger Clinic); U01HG006800 (Columbia University Health Sciences); U01HG6884 and U01HG006830 (Children’s Hospital of Philadelphia); U01HG6873, U01HG006388, and U01HG004609 (Northwestern University); U54MD007593 and U54MD007586 (Meheran Medical College); U01HG6876 (Partners Healthcare/Broad Institute); U01HG6864 (Baylor College of Medicine); U01HG06389 (Essential Institute of Rural Health, Marshall Field Clinic Research Foundation and Pennsylvania State University); U01HG06380 (Icahn School of Medicine at Mount Sinai); U01HG70, U01HG06385, and U01HG004603 (Vanderbilt University Medical Center serving as the Coordinating Center) from the NHGRI; funding to the eMERGE genotyping centers through grants U01HG004438 (CIRG) and U01HG004424 (the Broad Institute); funding for Vanderbilt University Medical Center’s Synthetic Derivative, Research Derivative, and BioVU by institutional funding and Clinical and Translational Science Award grant UL1RR000445 from National Center for Advancing Translational Sciences/NIH; by grant 2H200417-DI from the Intramural Research Program of the NHGRI; by funds from the Vanderbilt Ingram Cancer Center Ingram Professorship program (Dr. Wiener); and by grant 2H200417 from the Intramural Research Program of the National Human Genome Research Institute (Drs Zeng and Denny since joining the NIH).

Additional Contributions: Jacob Keaton, PhD, and Tracey Ferrara, PhD, NHGRI, NIH, provided logistical support. Jerome M. Visco, MD, Vanderbilt University Medical Center, Thomas Cassini, MD, NHGRI, NIH, and Huan Mo, MD, MD Anderson Cancer Center, provided critical comments. These individuals did not receive any compensation for this work.

Additional Information: Much of the work by Drs Zeng and Denny on this project was performed at Vanderbilt University before joining the NIH.

REFERENCES


© 2022 American Medical Association. All rights reserved.


