Research Article
Harnessing Population Pedigree Data and Machine Learning Methods to Identify Patterns of Familial Bladder Cancer Risk
Heidi A. Hanson\textsuperscript{1,2}, Claire L. Leiser\textsuperscript{1,3}, Brock O’Neil\textsuperscript{2}, Christopher Martin\textsuperscript{2}, Sumati Gupta\textsuperscript{4}, Ken R. Smith\textsuperscript{1,5}, Christopher Dechet\textsuperscript{5}, William T. Lowrance\textsuperscript{2}, Michael J. Madsen\textsuperscript{1}, and Nicola J. Camp\textsuperscript{1,6}

Affiliations:
\textsuperscript{1} Population Sciences, Huntsman Cancer Institute at the University of Utah, Salt Lake City, UT
\textsuperscript{2} Division of Urology, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT
\textsuperscript{3} Department of Epidemiology, University of Washington, Seattle, WA
\textsuperscript{4} Division of Oncology, Department of Medicine, Huntsman Cancer Institute at the University of Utah, Salt Lake City, UT
\textsuperscript{5} Department of Family and Consumer Studies, University of Utah, Salt Lake City, UT
\textsuperscript{6} Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT

Financial Support:
Research reported in this publication was supported by the National Institutes of Health K12 Award, 1K12HD085852-01; National Institutes of Health K07 Award 1K07CA230150-01, and HCI Cancer Center Support Grant (grant number P30CA042014). The Utah Cancer Registry is funded by the National Cancer Institute’s SEER Program, Contract No. HHSN261201800016I, the US Center for Disease Control and Prevention’s National Program of Cancer Registries, Cooperative Agreement No. NU58DP0063200, with additional support from the University of Utah and Huntsman Cancer Foundation.

Corresponding Author/Reprint Requests:
Heidi A Hanson, PhD, MS
Department of Surgery and Population Sciences, Huntsman Cancer Institute at the University of Utah, Salt Lake City, UT
Salt Lake City, UT 84132
Phone: 801-585-6794
Fax:
Email: heidi.hanson@hci.utah.edu

Conflicts of Interests:
The authors declare no potential conflicts of interest.

Running Title
Familial Bladder Cancer Phenotypes

Keywords
familial bladder cancer, urothelial cancer, family history of cancer, pan-cancer, multi-cancer risk

Word Count: 3286
Abstract: 250
Abstract

Background: Relatives of bladder cancer (BCa) patients have been shown to be at increased risk for kidney, lung, thyroid, and cervical cancer after correcting for smoking related behaviors that may concentrate in some families. We demonstrate a novel approach to simultaneously assess risks for multiple cancers to identify distinct multi-cancer configurations (multiple different cancer types that cluster in relatives) surrounding familial BCa patients.

Methods: This study takes advantage of a unique population-level data resource, the Utah Population Database (UPDB), containing vast genealogy and statewide cancer data. Familial risk is measured using Standardized Incidence Risk (SIR) ratios account for sex, age, birth-cohort, and person-years of the pedigree members.

Results: We identify 1,023 families with a significantly higher BCa rates than population controls (fBCa). Familial SIRs are then calculated across twenty-five cancer-types and a weighted Gower distance with K-medoids clustering is used to identify Familial Multi-Cancer Configurations (FMC). We find five FMCs, each exhibiting a different pattern of cancer aggregation. Of the 25 cancer types studied, kidney and prostate cancers were most commonly enriched in the familial BCa clusters. Laryngeal, lung, stomach, acute-lymphocytic leukemia, Hodgkin’s disease, soft tissue carcinoma, esophageal, breast, lung, uterine, thyroid, and melanoma cancers were the other cancer types with increased incidence in familial BCa families.

Conclusions: This study identified five familial BCa FMCs showing unique risk patterns for cancers of other organs, suggesting phenotypic heterogeneity familial BCa.

Impact: FMC configurations could permit better definitions of cancer phenotypes (subtypes or multi-cancer) for gene discovery and environmental risk factor studies.
Introduction

Bladder cancer (BCa) is the fifth most common cancer in the US with nearly 80,000 new cases per year.[1] Many genetic and environmental risk factors have been proposed, including smoking which is estimated to account for approximately 50% of cases and other chemical exposures such as aniline dyes. However, genetic causes remain largely unexplored. There is increasing evidence of familial clustering of BCa with first and second degree relatives of individuals with BCa having a 73% and 35% increase in risk of BCa, respectively.[2] This suggests an underlying germline genetic risk factor predisposing individuals to bladder cancer. Studies have tried to parse out the heritable and shared environmental components of risk using various methods and have estimated that somewhere between 7% and 12% of bladder cancers are due to heritable genetic risk factors and 12% due to shared environment.[3] Utilizing a longitudinal population based database, we show that combining big data analytics with pedigree and cancer registry data has the potential to identify and characterize ‘genetically-driven BCa subtypes’, ‘environmentally-driven BCa subtypes’, or ‘gene-environment BCa subtypes’.

Family history of cancer is an important risk factor for many cancers, which may extend across cancer types due to genetic pleiotropy or shared health behaviors [4-6]. For example, relatives of patients with BCa are at increased risk for not only bladder cancer, but also kidney, lung, thyroid, and cervical cancers [7-10]. In terms of familial multi-cancer configurations, these different etiological factors, genetic and environmental, may manifest as different multi-cancer configurations across a spectrum of organs. Breast cancer (BrCa) is an exemplar for genetically driven pleiotropic multi-cancer configurations in families. Approximately 30% of hereditary BrCa is explained by intermediate and high-risk inherited variants, like BRCA1 and BRCA2. Key factors contributing to the discovery of BRCA1 were dense familial clustering and coaggregation with ovarian cancer [11-13]. Unique multi-cancer configurations for carriers of BRCA1 mutations (breast, ovarian, Fanconi anemia, prostate, pancreatic,
fallopian tube and peritoneal cancers) and BRCA2 mutations (breast, male breast, prostate, pancreatic
cancers and Fanconi anemia) are now widely accepted (Figure 1). These multi-cancer configurations
were identified after discovery of BRCA1/2 mutations in BrCa. However, data driven methods make it
possible to uncover multi-cancer configurations before gene discovery. These configurations could
immediately permit better definitions of cancer phenotypes (subtypes or multi-cancer) for focused gene
discovery and environmental risk factor studies.

In contrast to BRCA1/2, smoking related cancers are an archetype for shared health behaviors
leading to familial multi-cancer configurations. Just as germline genetic mutations may lead to distinct
familial multi-cancer configurations, smoking related cancers may have a unique pattern of familial risk.
Parsing out whether multi-cancer configurations are related to gene, environment or a combination of
the two remains challenging for many cancers that have a strong environmental component.

BCa represents a clear example of this difficulty. While studies suggest a possible genetic link in
bladder cancer, teasing out this relationship remains a significant challenge because there is a strong
environmental etiological factor (e.g. smoking, chemical exposure). Previous studies suggest that
familial cancer risk in smokers may be the combination of gene and environmental exposures and
familial clustering of smoking related cancers has been demonstrated by multiple studies [7, 14].
Identifying families with multi-cancer configurations that appear to be related to shared environment,
shared genetics, or gene-environment interactions would allow us to categorize families into more
homogenous subtypes of cancer. Decreasing heterogeneity related noise in statistical analyses will
increase our ability to find meaningful genetic and environmental determinants of BCa and related
cancers.

Classical methods for assessing familial coaggregation are to determine the relative risk of
cancer in first-degree (FDR), second-degree (SDR), and third-degree (TDR) relatives of individuals with
the cancer of interest (pivot-individual). However, these methods use an iterative pairwise approach (pivot cancer and relative cancer) and are unable to simultaneously exploit data from multiple cancer types. As a result, power to identify novel patterns of multi-cancer configurations is limited. Other domains, such as marketing, have developed innovative computational techniques to discover patterns across expansive amounts of data in large databases and identify homogenous subpopulations of people. Application of such “big-data” techniques to linked genealogical and cancer databases can potentially identify new multi-cancer configurations and improve understanding of familial cancer risk, tumor spectrum and phenotypic heterogeneity.

This study takes advantage of a unique population-level data resource, the Utah Population Database (UPDB), containing vast genealogy and statewide cancer data. This study proposes an innovative method for identifying novel familial multi-cancer configurations for BCa through application of a network-inspired approach to complex family and cancer data.

**Methods**

**Study design and data**

We utilized the genealogical, demographic, and health data from the UPDB. The vast majority of individuals residing in Utah are represented in UPDB.[15-18] This immense genealogical data set is record-linked to many statewide datasets (including the Utah Cancer Registry), with annual updates. The full dataset contains nearly 5 M people with 28 M records and the infrastructure that links distinct records for a specific person allows the UPDB to create a depiction of the life history of an individual based on medical and administrative data. The UPDB supports hundreds of biodemographic, epidemiologic, and genetic studies primarily due to its comprehensive population coverage, pedigree complexity, and linkages across data sources [19, 20].
Cancer-specific data for individuals with urothelial BCa and their relatives was obtained from the Utah Cancer Registry (UCR), an original member of the Surveillance Epidemiology and End Results (SEER) program. We identified 6,752 individuals born 1900-1990 with urothelial BCa [2] and family history information (defined as at least twenty known relatives) as pivots. Families were only represented once in the analysis. When multiple siblings (n ≥ 2) had BCa (nsibsets = 104), one sibling was randomly selected as a pivot. Our sample included 6,416 three-generation families that included all first, second, and third degree relatives that were available in the UPDB.

Statistical Analysis

Familial Bladder Cancer

Our primary interest is familial BCa. For each family, familial risk for BCa (Supplementary Table 1) was measured using Standardized Incidence Risk (SIR) ratios accounting for the sex, age, birth-cohort, and person-years of the family members (for a detailed description of SIR calculations, see supplemental methods). Person-years were calculated using the minimum of the first year residing in Utah or 1966 to the year of first cancer diagnosis, last year of residence in Utah (due to death or migration), or 2017. Further analyses were restricted to the sample to families with a statistical significant BCa SIR and more than one case of BCa in the family. This resulted in a final sample size of 1,023 familial BCa families for the multi-cancer configuration analysis.

Familial Multi-Cancer Enrichment

As for BCa above, familial risk for twenty-five additional cancers were measured using the SIR accounting for the sex, age, birth-cohort, and person-years of the family members (Supplementary Table 1). Two risk metrics were used to capture the family’s multi-cancer signature. First, wSIR, which is the SIR weighted by the p-value, to incorporate both the magnitude and significance of the familial risk. This
was calculated using the following equation, and allows us to include, but down weight, SIR values that were not significantly different relative to the overall population.

\[ wSIR_{ij} = SIR_{ij} \cdot (1 - \log(p_{ij})) \]

Where \( p \) is the \( p \)-value, \( i \) is the family, and \( j \) is the cancer type.

In order to avoid bias due to large SIRs (especially for rare cancers), and increase the robustness of our results, we imposed a maximum value such that any \( wSIR \) values larger than the 90\(^{th}\) percentile were set to the 90\(^{th}\) percentile value across all families for the cancer type.

\[ wSIR \begin{cases} wSIR_{ij}; & \text{if } wSIR_{ij} \leq wSIR_{j90} \\ wSIR_{j90}; & \text{otherwise} \end{cases} \]

where 90 indicates the 90\(^{th}\) percentile for cancer \( j \).

Second, we considered the significance of a SIR as a dichotomous indicator of risk (\( I_{SIR} \)). Families were considered to have “high risk” status (\( I_{SIR} = 1 \)) for a cancer type if the SIR was statistically greater (\( p < 0.05 \)) than the age and sex adjusted rates for the Utah population and “population risk” (\( I_{SIR} = 0 \)) otherwise. As all families were at significantly increased risk of BCa by design, we did not include an \( I_{SIR} \) matrix for BCa (i.e., we had 26 measures of \( wSIR \) and 25 measures of \( I_{SIR}; p = 51 \)).

**Familial Multi-Cancer Configurations**

We constructed a 1,023 x 51 matrix (families x \( (wSIR, I_{SIR}) \)) and calculated the Gower general coefficient (\( daisy \) function in the \textit{cluster} package in R\[21\]) to measure similarities between multi-cancer configurations of families\[22, 23\]. The Gower distance was selected because our data had both continuous and categorical (indicator) variables and it allows for mixed variables to be used simultaneously (detailed information can be found in the supplement). We used partitioning around medoids (PAM or K-medoids clustering package in R\[21\]) to measure similarities between the multi-cancer risk signatures of families. \( K \) was selected by running a series of iterative models for \( k=2 \) to \( k=20 \).
and using Silhouette plots to identify the point of diminishing improvement in average Silhouette width.

The Silhouette plot for the final model, K=5, is displayed in Supplementary Figure 1.

Bootstrapping was used to evaluate the reproducibility of the clustering using the `clustboot` function in R. The PAM algorithm was used with 200 random draws and the results from each draw were stored in a results matrix, transformed into a consensus matrix using the ward linkage algorithm and the `consensusmatrix` function in R, and then plotted in a heatmap for visualization. The results for k=5 were fairly stable, with some switching between clusters 3 and 5 (Supplementary Figure 2).

Assessment of Families by Multi-Cancer Configuration

Fixed effect meta-analysis using the inverse variance method for pooling was used to estimate the cluster specific differences in SIRs and their 95% CIs using the R package `metagen`. Known profiles of risk were used to test the FMCs for risk in smoking related cancers (lung, mouth, lips, nose and sinuses, larynx, pharynx, esophagus, stomach, pancreas, kidney, uterus, cervix, colon/rectum, ovary, and acute myeloid leukemia), Lynch syndrome cancers (small intestine, colon, pancreas, uterus, and kidney/renal pelvis), and arsenic related cancers (lung, prostate, and kidney; non-melanoma skin was not included because it is not reported to cancer registries).

Results

Familial BCa families were classified as those with a significantly increased risk of BCa relative to the general population and the family had more than one BCa. We identified 1,023 familial BCa families, each centered around a BCa pivot individual. These 1,023 familial BCa (fBCa) pivots had 59,177 relatives and 829 spouses, with the number of family members ranging from 16 to 628 (Table 1). Median age at diagnosis for all fBCa patients was 72.3 and ranged from 15 – 98. Median age at diagnosis of fBCa pivots was slightly earlier (71.19; p<0.001) and ranged from 27 to 96 years. In the overall sample, 19.2% of the
BCa cases were female. When stratified by familial BCa, we found that a higher proportion of fBCa pivots were female relative to non-fBCa cases (20.6% vs. 18.2%; p=0.03).

**Familial Multi-Cancer Configurations (FMCs)**

Using the 1,023 familial BCa families, we found five distinct FMCs (FMC1-5) using k-medoid clustering; each exhibiting a different patterns of cancer aggregation or multiphenogram (Figure 2). The proportion of BCa families captured by the five fBCa FMCs were 25.3%, 16.7%, 30%, 24.2%, and 3.8%, respectively. Of the 25 cancer types studied, kidney and prostate were most commonly enriched in the familial BCa multi-cancer configurations. Laryngeal, lung, stomach, acute-lymphocytic leukemia, Hodgkin’s disease, soft tissue carcinoma, esophageal, breast lung, uterine thyroid, and melanoma were the other cancer types found to have increased incidence in familial BCa families. The clustering algorithm accounted for both the magnitude of the SIR and whether or not there is a statistically significant increased risk for the cancer type (wSIR and I_{SIR}). Similarly, both factors were considered when characterizing each FMC, or FMC signature (Figure 2). An FMC was defined as “strongly enriched” for cancer types that were both significantly increased in at least 10% of families within the FMC (Figure 3) and for which the SIR for the cluster (combination across all families in the FMC) was statistically significant (Figure 4 and Supplementary Table 2).

All families had a minimum of two BCa diagnoses by definition and a maximum of seven, but FMCs varied in the magnitude of their BCa risk. The mean BCa SIR across all 1,023 families was 12.12 (median of 9.52). There was significant heterogeneity in the average BCa SIR across FMCs (Figure 5 and Supplementary Table 2), with average BCa SIR ranging from 5.31 (FMC2) to 19.98 (FMC1).

In addition to extremely high risk for BCa, FMC1 was strongly enriched for both kidney and prostate cancer. The average SIR for these cancers was 1.21 (95% CI 1.10, 1.34; 10% families) and 1.73 (95% CI 1.57, 1.90; 10% families), respectively (Supplementary Table 2). FMC2 was strongly enriched for
larynx (SIR=5.97, 95% CI 5.22, 6.83; 37% of families), stomach (1.19, 95% CI 1.06, 1.34; 11% of families), and lung cancer (1.66, 95% CI 1.47, 1.86; 11% of families). FMC3 did not have an increased risk for any other cancer type other than BCa. FMC4 was strongly enriched for prostate, kidney and renal pelvis, and Acute Lymphocytic Leukemia (ALL), with average SIRs of 1.72 (95% CI 1.57, 1.88; 11% of families), 1.28 (95% CI 1.16, 1.41; 11% of families), and 1.25 (95% CI 1.13, 1.37; 11% of families). FMC5 was strongly enriched for uterine, prostate, thyroid, melanoma, kidney, small intestinal, soft tissue, and Hodgkin’s lymphoma. Small intestinal cancer had the highest risk, with nearly all families having a significant increased risk of this rare cancer SIR 29.67 (95% CI 21.84, 40.29; 97% of families). The significance for small intestine cancers in almost all families is striking. Risk for the other cancer sites in FMC5 ranged from 1.37 (Hodgkin’s disease; 16% of families) to 1.88 (Melanoma; 11% of families).

**Smoking and Lynch Syndrome multi-cancer Profiles**

We calculated combined SIRs for groups of cancers related to: smoking, lynch syndrome, and arsenic exposure using fixed effect meta-analysis and inverse variance method. While all FMCs were at increased risk for arsenic related cancers related to the population controls, there was not a significant difference in risk across the FMCs. However, SIR estimates for smoking related cancer and lynch syndrome related cancers varied by FMC (Figure 6 panel A&B and Supplementary Table 3). FMC2 had the highest smoking related SIR (SIR\textsubscript{smoking} = 1.44; 95% CI 1.39, 1.49), followed by FMC5 (SIR\textsubscript{smoking} = 1.36; 95% CI 1.26, 1.47). Both of these are significantly higher than the lowest: FMC1 (SIR\textsubscript{smoking} = 1.17; 95% CI 1.14, 1.21) and FMC4 (SIR\textsubscript{smoking} = 1.19; 95% CI 1.15, 1.22) and FMC3’s was only slightly higher (SIR 1.23; 95% CI 1.20, 1.27). Even after excluding small intestinal cancer (Figure 6 panel C), the combined SIRs for cancers related to lynch syndrome also varied by FMC. The FMC5 cancer risk profile most resembled Lynch syndrome, and showed a nearly two-fold increase in lynch related cancers (SIR =1.96; 95% CI 1.78, 2.16). The risk in FMC5 remained the highest after accounting for small intestine cancers.
Decreased risk for some cancer types

In addition to a significant increased cancer risk, some FMCs were characterized by having zero enrichment for certain cancer types. Specifically, no families in FMC1, FMC3, and FMC4 had increased risk for larynx or small intestinal cancers, which may suggest these FMCs are not related to smoking.

Discussion

We have described a method for identifying familial multi-cancer configurations (FMC), and illustrated 5 distinct patterns of multi-cancer risk surrounding familial BCa patients. The five different FMC clusters identified for familial BCa illustrate the potential of our network-inspired approach to simultaneously assess multiple cancer risks, shifting focus away from a unidimensional definition of family history to a more comprehensive view of family history and risk identification. The pattern of multi-cancer clustering in FMC2 suggests that cancer risk in those families is driven by smoking and other related exposures, while the pattern in FMC5 is similar to patterns of cancer clustering in Lynch families. FMC1, FMC3, and FMC4 have patterns suggest shared genetics or environments other than smoking may play a strong role in cancer risk in these families. We did not find a single FMC pattern consistent with arsenic exposure.

There are limitations with our method of familial multi-cancer clustering discovery. First, this method did not factor in age at diagnosis or histopathological information. Future versions of this method would be strengthened by considering those factors. Third, this method did not utilize all information in pedigree structure, such as weighting by kinship coefficient, and future studies should test methods that take advantage of that information. Despite these limitations, application of this approach could provide important insight to numerous cancers and other age-related chronic diseases. Future work will also need to investigate whether there are identifiable gene, environment or
gene/environment factors that explain the observed differences in FMCs in a similar way as BRAC1/2 result in distinct multi-cancer clusters.

The concept of phenomes, extending the phenotype to include multiple disease types to characterize genotype-phenotype relationships, is not new. However, the phenome is usually referred to as the set of all phenotypes within an individual. Recent studies demonstrated the feasibility of ‘phenome-wide association scans’ (PheWAS), using genetic data linked to medical records to identify multiple phenotypes associated with a single genotype [24]. Pan-cancer analyses are another familiar concept. The Cancer Genome Atlas (TCGA) launched the Pan-cancer analysis project in 2012 with the goal of combining molecular data across large numbers of tumor types to compare -omics data across tumors, potentially allowing for the identification of thematic pathways [25]. However, these are independent individuals. Here we utilized family-based data to investigate familial patterns of pan-cancer clustering, a novel familial extension to pan-cancer studies.

Molecular diagnostics and cancer-specific subtypes are becoming an essential component of clinical decision making, however current approaches are typically organ-specific or do not include context for understanding the interplay between genes and environment. The most recent BCa-specific analysis of TCGA data found five distinct subtypes; 1) luminal-papillary, 2) luminal-infiltrated, 3) luminal, 4) basal-squamous, and 5) neuronal [26]. Other studies have shown these subtypes closely align to cancers in other organs. The PAM50 algorithm (originally developed to classify breast cancers) has also been used to classify prostate, bladder and lung cancer [27-30], evidence that molecular subtypes may share common etiologies. Subtypes of BCa cancer have clinically meaningful differences [31], however what predisposes an individual to a particular subtype remains unknown. Understanding etiology of the disease has clinical potential because it allows for increased screening in at-risk populations and may guide treatment decisions at early stages of diagnosis. Future work to combine familial multi-cancer
phenotypes, as we developed here, with tumor molecular data has potential for identification and characterization of BCa subtypes that may share common etiological and/or tumorigenic pathways.

Many genetic and environmental risk factors have been proposed in BCa risk, which could manifest as different multi-cancer configurations across a spectrum of organs. FDRs likely share similar environments throughout the life course and therefore familial aggregation of BCa cancer may not be entirely genetic. For example, the familial multi-cancer configuration pattern in FMC2 appears to be strongly related to smoking related exposure. Genetic predispositions may also make individuals sensitive to environmental exposures. For example, arsenic metabolism may vary between individuals and the rate of metabolism affects risk for adverse health outcomes [32]. Moreover, individuals with N-acetyltransferase 2 (NAT2) slow acetylator and glutathione S-transferase μ1 (GSTM1)-null genotypes may have increased risk for BCa when exposed to carcinogens through smoking or occupational risk [33-35].

Conclusions

This study identified five familial BCa FMCs. These vary by BCa risk and by the age at diagnosis and sex of the pivot BCa case. Additionally, each FMC shows unique risk patterns for cancers of other organs. Leveraging genealogical data linked to cancer records is a powerful way to perform cross-phenotype, multi-cancer analyses.
References


<table>
<thead>
<tr>
<th>Table 1. Number of Family Members and Cancer Diagnoses by Familial Multi-Cancer Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FMC 1 (N=234)</strong></td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td><strong>Min</strong></td>
</tr>
<tr>
<td>Number of Family Members</td>
</tr>
<tr>
<td>Acute Lymphocytic Leukemia</td>
</tr>
<tr>
<td>Acute Myeloid Leukemia</td>
</tr>
<tr>
<td>Bladder</td>
</tr>
<tr>
<td>Brain</td>
</tr>
<tr>
<td>Breast</td>
</tr>
<tr>
<td>Central Nervous System</td>
</tr>
<tr>
<td>Cervical</td>
</tr>
<tr>
<td>Colon</td>
</tr>
<tr>
<td>Esophagus</td>
</tr>
<tr>
<td>Hodgkin-Nodal</td>
</tr>
<tr>
<td>Kidney</td>
</tr>
<tr>
<td>Larynx</td>
</tr>
<tr>
<td>Liver</td>
</tr>
<tr>
<td>Lung and Bronchus</td>
</tr>
<tr>
<td>Melanoma</td>
</tr>
<tr>
<td>Myeloma</td>
</tr>
<tr>
<td>Non-Hodgkin Lymphoma</td>
</tr>
<tr>
<td>Ovary</td>
</tr>
<tr>
<td>Pancreas</td>
</tr>
<tr>
<td>Prostate</td>
</tr>
<tr>
<td>Small Intestine</td>
</tr>
<tr>
<td>Soft Tissue</td>
</tr>
<tr>
<td>Stomach</td>
</tr>
<tr>
<td>Testis</td>
</tr>
<tr>
<td>Thyroid</td>
</tr>
<tr>
<td>Uterine</td>
</tr>
</tbody>
</table>
Figure 1. Example of a familial multiphenogram for known genetic mutations. BRCA1 (Panel A) and BRCA2 (Panel B). Familial multiphenograms illustrate patterns of familial multi-cancer clustering (FMC) that are unique to the underlying etiological sub-type of breast cancer. Combining machine learning tools with large, familial databases can enable the identification of these unique patterns and allow for sub-classification of tumors into more homogenous subtypes.

Figure 2. Familial multiphenograms for familial bladder cancer families. Five distinct familial multi-cancer clusters (FMC1 – FMC5) were discovered and ranged in size from 37 families (FMC5) to 289 families (FMC3) that had similar patterns of familial cancer clustering. Multiphenograms are labeled with their cancer type and the percent of the families in the FMC with statistically significant increased risk for the respective cancer type.

Figure 3. Proportion of families with significantly increased risk of cancer relative to the general population by familial multi-cancer cluster (FMC). Five distinct FMCs were discovered (FMC1 – FMC5) and ranged in size from 37 families (FMC5) to 289 families (FMC3). Families in each FMC had a similar pattern of familial cancer clustering. While not all families were at significant increased risk for each cancer displayed, the magnitude of risk was similar within each FMC.

Figure 4. Standardized Incidence Risk (SIR) profile for each familial multi-cancer cluster (FMC). Displayed SIRs are based on fixed effect meta-analysis using the inverse variance method for pooling was used to estimate the cluster specific differences in SIRs. SIRs significant at p<0.05 are displayed in panels FMC1 – FMC5.

Figure 5. Bladder cancer Standardized Incidence Risk (SIR) profile by familial multi-cancer cluster (FMC1 – FMC5). All families are at statistically significantly increased risk of bladder cancer (BCa) by definition, however there is variation in the magnitude of the effect. Displayed SIRs are based on fixed effect meta-analysis using the inverse variance method for pooling was used to estimate the cluster specific differences in SIRs.

Figure 6. Standardized Incidence Risk (SIR) profile by familial multi-cancer cluster (FMC1 – FMC5). Panel A. Estimated SIR for smoking related cancer by FMC. Panel B. Estimated SIR for Lynch syndrome cancers by FMC. Panel C. Estimated SIR for Lynch syndrome cancers by FMC. See supplemental material for a list of cancers included in the smoking and Lynch syndrome designations.
Figure 1

A. Familial cancer configuration 1: BRCA1 carriers with Breast cancer as the cancer of interest

B. Familial cancer configuration 2: BRCA2 carriers with Breast cancer as the cancer of interest
Figure 2
Figure 3

No additional cancers on this risk profile
Figure 6
Cancer Epidemiology, Biomarkers & Prevention

Harnessing Population Pedigree Data and Machine Learning Methods to Identify Patterns of Familial Bladder Cancer Risk

Heidi A Hanson, Claire L Leiser, Brock O'Neil, et al.

Cancer Epidemiol Biomarkers Prev  Published OnlineFirst February 25, 2020.

Updated version
Access the most recent version of this article at:
doi:10.1158/1055-9965.EPI-19-0681

Supplementary Material
Access the most recent supplemental material at:
http://cebp.aacrjournals.org/content/suppl/2020/02/25/1055-9965.EPI-19-0681.DC1

Author Manuscript
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, use this link http://cebp.aacrjournals.org/content/early/2020/02/25/1055-9965.EPI-19-0681. Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.