Analysis of Co-aggregation of Cancer Based on Registry Data

Abigail G. Matthews, BA[1,2]

Rebecca A. Betensky, PhD[1,2]

Hoda Anton-Culver, PhD[3]

Deborah Bowen, PhD[4]

Constance Griffin, MD[5]

Claudine Isaacs, MD[6]

Carol Kasten, MD[7]

Geraldine Mineau, PhD[8]

Susan Nayfield, MD, Msc[7]

Joellen Schildkraut, PhD[9]

Louise Strong, MD[10]

Barbara Weber, MD[11]

Dianne M. Finkelstein, PhD[1,2]

[1] Massachusetts General Hospital Biostatistics Center, Boston, MA

[2] Department of Biostatistics, Harvard School of Public Health, Boston, MA

[3] University of California, Irvine, CA

[4] Fred Hutchinson Cancer Research Center, Seattle, WA

[5] Johns Hopkins University, Baltimore, MD

[6] Georgetown University Lombardi Cancer Center, Washington DC

[7] National Institutes of Health, Bethesda, MD

[8] University of Utah, Salt Lake City, UT

[9] Duke University Medical Center, Durham, NC

[10] University of Texas MD Anderson Cancer Center, Houston, TX

[11] University of Pennsylvania, Philadelphia, PA

Address for correspondence: Dianne M. Finkelstein, PhD, Massachusetts General Hospital Biostatistics Center, 50 Staniford Street, Suite 560, Boston, MA 02114; Fax: (617) 724-9878; E-mail: dfinkelstein@partners.org

Abstract

Objective. Conduct an exploratory analysis of co-aggregation of cancers using registry-based data. **Methods.** We utilized sibships from over 18,000 families who had been recruited to the NCI-sponsored multi-institutional Cancer Genetics Network. The analysis evaluates co-aggregation at the individual- and family-level and adjusts for ascertainment. **Results.** We found statistically significant familial co-aggregation of lung cancer with pancreatic ($p<0.0001$), prostate ($p < 0.001$), and colorectal cancers ($p=0.003$). In addition, we found significant familial co-aggregation of pancreatic and colorectal cancers ($p=0.022$), and co-aggregation of hematopoietic and (non-ovarian) gynecologic cancers ($p=0.01$). **Conclusion.** This analysis identified novel associations between lung cancer and several other GI and GU cancers.

Key Words: **Familial aggregation, Association, Family study, Cancer Genetics Network**

Introduction

The Cancer Genetics Network (CGN) is a multi-institutional consortium that was developed by the National Cancer Institute as a resource for epidemiological and translational research into the genetic basis of cancer susceptibility. Since 1999, over 18,000 individuals (probands) have consented to participate in the Core Registry, and complete family and medical history has been collected on each participant and resides in a Core Registry. This Core database is maintained and updated regularly both to retain contact and communication with CGN participants to invite them to participate in translational research (such as cancer screening and psychosocial research) studies, and to provide a resource for hypothesis-generating database studies of the genetic basis of cancer. One natural study that can be readily performed utilizing such registry data that includes family cancer history is an analysis of co-aggregation of disease in individuals and families. Such an identification of cancers that co-aggregate can be useful for understanding the etiology of disease. In addition, this knowledge can lead to a more focused screening for earlier detection of disease, often resulting in improved survival.

There have been many reports in the literature on evidence of cancers that aggregate in families. Reviews of the literature on familial aggregation of breast, ovarian and colorectal cancers are given in Hoffman et al. [1], Berchuck [2], and Bonaïti-Pellié [3] respectively. Narod [4] reported that prostate cancer also aggregates within families. More recent literature has reported on familial aggregation of pancreatic [5], hematopoietic [6], and lung cancers [7]. Co-aggregation of pairs of distinct cancers has also been reported in the literature: colorectal cancer is known to co-aggregate with breast and ovarian cancers [8-10], and several studies have shown that breast and ovarian cancers cluster within families and within individuals [11,12], primarily due to mutations in BRCA1/2 [13]. Studies have also suggested that breast and ovarian cancers each co-aggregate with other gynecologic cancers, but none of these results was statistically significant [10,14].

Studies of co-aggregation of multiple less prevalent cancers require a large database of family medical history of disease such as the Cancer Genetics Network has developed. This paper is a report of the experience and results of an analysis of the CGN Core Registry that was undertaken to explore for novel evidence of cancers that co-aggregate at the individual- and family-level.

Materials and Methods

Study Population

The Cancer Genetics Network is a multi-site NCI-sponsored research consortium that recruits participants at each of eight clinical sites. Recruitment is population-based at some institutions (11,628 families) and is based on clinic-, physician- or self-referral at others (6,253 families). Once a proband is recruited, information is collected on all cancer diagnoses of the proband's first, second and third degree relatives. The disease statuses of the probands are confirmed, but those of their family members are not. Anton-Culver et al. [15] gives a detailed description of the CGN Registry, and of the specific ascertainment schemes that were utilized.

For the purpose of this analysis, disease sites were combined into categories as in DeVita et al. [16]: breast (female cases only), ovarian, prostate, colorectal, non-ovarian gynecologic, pancreatic, hematopoietic (primarily bone marrow) and lung. Gynecologic cancers consist mainly of cervical and uterine/endometrial cancers. Males were included in the single disease analyses of non-gender-specific cancers and analyses involving prostate cancer. Similarly, women were excluded from any analysis involving prostate cancer. The CGN participants analyzed in this paper consist of over 65,000 siblings (including all probands) who were recruited prior to January 2003.

Statistical Methods

For the analysis of multiple cancers, it is necessary to choose a method that appropriately captures the association between diseases and adequately adjusts for ascertainment. Thus, it would not be appropriate to use simple odds-ratio calculations to identify cancers that cluster in families because this approach does not adjust for the co-aggregation of both diseases when assessing the degree of aggregation of each disease individually. For example, a simple odds ratio approach is not able to address whether ovarian cancer aggregates in families above and beyond its co-aggregation with breast cancer.

Hudson et al. [17] proposed a *family predictive model* that provides a method to adjust for all possible relationships between two diseases within families and within individuals. In addition, this method appropriately adjusts for the fact that some families are not population-based. The family predictive model specifies the log-odds of disease as a linear function of the number of relatives with disease. Familial aggregation is tested by assessing the departure of the regression coefficient from zero. This model can be extended to include individual covariates and pair-level predictors. This model is not generally applicable to varying family sizes [18,19], and thus we restricted the analyses to sibships consisting of between two and five members. Only sibships were used in order to ensure approximate environment and age matching. For the analysis of aggregation of the female- (male-) specific cancers only sisterhoods (brotherhoods) were used. The model for aggregation of lung cancer included a covariate indicating whether the proband had ever smoked.

The analysis of multiple distinct cancers (say, *A* and *B)* that co-aggregate in families used the *multivariate* family predictive model of Hudson et al. [17]. The simplest form of the model specifies the log-odds of disease *A* as a linear function of an individual's disease *A* status, the number of their siblings with disease *A*, and the number of their siblings with disease *B*. For example, the log-odds of lung cancer for an individual is a linear function of his/her colorectal cancer status, the number of siblings with lung cancer, and the number of siblings with colorectal cancer. The coefficients of the model used for this analysis capture: (i) co-aggregation of colorectal and lung cancers within *individuals*, (ii) aggregation of lung cancer within families; (iii) aggregation of colorectal cancer in families; and (iv) co-aggregation of colorectal and lung cancers within *families*.

Logistic regression underlies the family predictive model. Let $y_{k,j}$ denote the disease k ($k=A,B$) status of the jth individual, $s_{k,-j}$ denote the number of their siblings with disease k, and p_k denote the probability of disease k conditional on all other cancer outcomes in the family. Then, the simplest multivariate family predictive model implies the following logistic regression equations for the conditional log-odds of each disease:

$$\mathrm{logit}[p_A(j)] \quad = \quad a_A + \delta\, y_{B,j} + \gamma_A\, s_{A,-j} + \gamma_{AB}\;\; s_{B,-j}$$

$$\mathrm{logit}[p_B(j)] \quad = \quad a_B + \delta\, y_{A,j} + \gamma_B\, s_{B,-j} + \gamma_{AB}\, s_{A,-j} \quad . \qquad (1)$$

The parameters in this model have conditional interpretations: a_A (a_B) is the log-odds of disease *A (B*) given no other disease *A (B*) in the family, δ is the log-odds ratio for co-aggregation of diseases *A* and *B* within individuals, γ_{AB} is the log-odds ratio for co-aggregation of diseases *A* and *B* between family members, and γ_A (γ_B) is the log-odds ratio for aggregation of disease *A (B*). We note that the estimates of both levels of co-aggregation derived from the model are not useful because this basic application of the family predictive model treats the diseases as exchangeable with respect to co-aggregation. For example, at the individual (as well as family) level, the increase in the risk of lung cancer associated with having colorectal cancer is assumed to be of the same magnitude as the increase in the risk of colorectal cancer associated with having lung cancer. Although this simplifying assumption may not be valid for all diseases, especially in the case of uncommon diseases, the data are typically too sparse for a more complex model. Although the parameter estimates from these analyses may not be appropriate for prediction, they do form the basis for valid tests of association and thus we will focus only on the statistical inference about the co-aggregation of cancers that is provided by these methods.

The CGN Registry includes families recruited due to a personal or family history of cancer. To account for this ascertainment, we treated the proband's disease status as fixed by design. Thus probands enter our logistic regression models only as covariates and not as outcomes; they contribute to the number of relatives with disease.

The machinery of generalized estimating equations (GEEs) [20] is used to adjust for the correlation among family members. A two-sided significance level of

0.05 was used in all tests. The p-values are reported without adjustment for multiple comparisons as this is viewed as an exploratory analysis.

Results

There were 12,263 families used in this analysis. There were 1,159 colorectal cancers, 450 lung cancers, 185 hematopoietic cancers, and 149 pancreatic cancers. For female cancers, our analysis was based on 9,749 sisterhoods containing 5,972 cases of breast cancer, 677 of ovarian cancer and 571 cases of non-ovarian gynecologic cancer. For prostate cancer, the analysis was based on 8,072 brotherhoods with 3,264 cases of prostate cancer.

Familial Co-aggregation of Individual Cancers

Evaluation of familial aggregation of a single disease is driven by the number of families with two or more cases of disease. Table 1 gives the distribution of the number of cancers within sibships as well as the results of the family predictive models for each cancer individually. Our results confirmed the results published in earlier papers reporting familial aggregation of breast cancer [1], ovarian cancer [2], colorectal cancer [3], prostate cancer [4], pancreatic cancer [5], hematopoietic cancer [6], and lung cancer [7].

Familial Co-aggregation of Distinct Cancers

In considering familial co-aggregation of two distinct cancers within families and within individuals, the number of families and individuals with at least one case of each disease drives co-aggregation. Table 2 gives the number of individuals with two (or more) cancers in a pair-wise fashion as well as the p-values from the multivariate family predictive models assessing co-aggregation of cancers. These results for co-aggregation at the family-level are given in Table 3. Cancer sites that were too rare and those which have already been reported as co-aggregating are not listed in these tables but results are available from the authors. Note that these models are different from those in Table 1 in which only one cancer was considered at a time.

Our results confirmed those published in earlier papers reporting co-aggregation of breast and ovarian cancers [11,12,14], and co-aggregation of colorectal and prostate cancers [5]. In addition, we identified novel associations. We found that lung cancer co-aggregates with pancreatic cancer (p<0.0001), prostate cancer (p<0.001), and colorectal cancer (p=0.003). Other novel results include the finding that hematopoietic and non-ovarian gynecologic cancers cluster together at the family-level (p=0.011). Also, pancreatic cancer co-aggregates in families with colorectal cancer (p=0.022). At the individual-level, both hematopoietic and lung cancers co-aggregate *negatively* with breast cancer (p=0.027 and p=0.030, respectively).

Discussion

The analysis of the CGN Registry provided interesting results on cancer sites, such as lung cancer, for which the hereditary form of the disease is believed to be quite rare in the general population [21], and hematopoietic cancer, for which the location of the responsible gene or genes is unknown [22,23]. It would be useful to try to further study the genetic and/or environmental factors responsible for the familial clustering of these cancers.

There are several limitations to studying disease aggregation using data collected in a family registry such as the CGN. First, there are issues with misreporting of disease history. The disease statuses of probands were confirmed by the CGN sites, but not that of their family members. We note that reporting errors can occur both because many deep organ cancers are not accurately recalled by the proband, and also metastatic sites are sometimes reported as primary cancer sites [24,25]. Second, information on the behavioral history (such as smoking history) of relatives was not recorded, so we were only able to adjust for the proband's smoking status. This information would have been useful for the analysis of smoking related cancers such as lung cancer. In absence of this, the proband's smoking status must be viewed as surrogate information. Third, our analyses assumed that all families were sampled because of the disease status of the proband. In actuality, the ascertainment was more complex. For example, some probands referred themselves to the Network. It is not known why these probands chose to participate; it may be due to a family (not personal) history of cancer. In this case a familial association may be induced solely from the ascertainment scheme, and should be appropriately included in the analysis. Our current research focuses on evaluating familial aggregation studies with such complex ascertainment schemes. Lastly, evaluation of co-aggregation of cancers within individuals is complicated by competing risks [26], that is, an individual may die of lung cancer before developing another cancer. This is especially problematic when dealing with cancers with high mortality rates, such as ovarian, lung, pancreatic and hematopoietic cancers [27]. This would tend to diminish the evidence of aggregation. One approach to decrease the effects of competing risks is to adjust for age. This would also adjust for individuals who never had cancer before study participation including those who died before developing disease. Our current research also focuses on developing methods of applying the family predictive model to account for the ages at disease onset and censoring.

Despite these limitations, our analysis revealed several interesting disease clusterings, which could be useful in guiding future research into the genes and environmental factors associated with cancer susceptibility. The CGN resource of carefully collected family history of cancer and consent for future contact for research studies in these diseases should be viewed as a rich source available to the scientific community for cancer genetics research. Further information on this resource is available on the web [28,29].

Table 1. Aggregation of individual cancers

Cancer	Cases per sibship	No. of sibships (%)	p-value
Hematopoietic	0	12,082 (98.5)	**0.003**
	1	177 (1.4)	
	2	4 (0.0)	
	3 or more	0 (0.0)	
Lung	0	11,849 (96.6)	**<0.0001**
	1	381 (3.1)	
	2	30 (0.2)	
	3 or more	0 (0.0)	
Pancreatic	0	12,116 (98.8)	**0.104**
	1	145 (1.2)	
	2	2 (0.0)	
	3	0 (0.0)	
Prostate*	0	5,505 (68.2)	**<0.0001**
	1	2,006 (24.9)	
	2	456 (5.7)	
	3 or more	105 (1.3)	
Colorectal	0	11,179 (91.2)	**<0.0001**
	1	1,014 (8.3)	
	2	66 (0.5)	
	3 or more	4 (0.0)	
Breast*	0	4,871 (50.0)	**<0.0001**
	1	3,911 (40.0)	
	2	853 (8.8)	
	3 or more	114 (1.2)	
Ovarian*	0	9,093 (93.3)	**0.023**
	1	636 (6.5)	
	2	19 (0.2)	
	3 or more	1 (0.0)	
Gynecologic*	0	9,212 (94.5)	**<0.0001**
(non-ovarian)	1	506 (5.2)	
	2	28 (0.3)	
	3 or more	3 (0.0)	

* Single-sex sibships only.

Table 2. Co-aggregation of two different cancers within an individual

# individuals							
with both							
Cancers							
(p-value)	**Lung***	**Pancreatic***	**GYN****	**Prostate*****	**Breast****	**Ovarian****	**Colorectal***
Hematopoietic*	1 (p=0.76)	0 †	5 **(p=0.011)**	14 (p=0.09)	9 **(p=0.027)**	3 (p=0.83)	11 (p=0.77)
Lung*	-	2 (p=0.38)	7 (p=0.34)	23 (p=0.05)	35 **(p=0.030)**	8 (p=0.37)	20 (p=0.78)
Pancreatic*	-	-	2 †	1 †	2 (p=0.13)	1 †	1 (p=0.93)
Gynecologic**	-	-	-	-	150 (p=0.41)	31 **(p<0.0001)**	7 (p=0.08)
Prostate***	-	-	-	-			100 (p=0.65)
Breast**	-	-	-	-	-	151 **(p<0.0001)**	104 (p=0.56)
Ovarian**	-	-	-	-	-	-	12 (p=0.54)

*39,572 individuals

** 27,334 females

*** 22,300 males

† GEEs cannot be performed – cancers are too rare

Table 3. Co-aggregation of multiple cancers within a sibship

% families with at least one of each disease **(p-value)**	**Lung**	**Pancreatic**	**GYN***	**Prostate**	**Breast**	**Ovarian**	**Colorectal**
Hematopoietic	0.1% (p=0.8))	0.0% †	0.1% **(p=0.01)**	0.9% (p=0.67)	0.4% (p=0.48)	0.1% (p=0.87)	0.2% (p=0.84)
Lung	-	0.1% **(p<0.0001)**	0.2% (p=0.50)	2.4% **(p<0.001)**	1.3% (p=0.51)	0.3% (p=0.44)	0.5% **(p=0.003)**
Pancreatic	-	-	0.1% †	0.5% †	0.2% (p=0.14)	0.0% †	0.1% **(p=0.022)**
Gynecologic*	-	-	-	-	3.7% (p=0.21)	0.6% (p=0.86)	0.9% **(p=0.002)**
Prostate	-	-	-	-	-	--	4.2% **(p=0.004)**
Breast	-	-	-	-	-	3.8% **(p=0.034)**	2.5% **(p<0.001)**
Ovarian	-					-	0.5% (0.11)

* Non-ovarian.

† GEEs cannot be performed – cancers are too rare

References

1. Hofmann W, Schlag PM. BCRA1 and BRCA2 – breast cancer susceptibility genes. J Cancer Res Clin Oncol 2000; 126:487-496.

2. Berchuck A, Carney M, Lancaster JM, Marks J, Fitreal AP. Familial breast-ovarian cancer syndromes: BRCA1 and BRCA2. Clin Obstet Gynecol 1998;41:157-166.

3. Bonaïti-Pellié C. Genetic risk factors in colorectal cancer. Eur J Cancer Prev 1999;8 Suppl 1:27-32.

4. Narod S. Genetic epidemiology of prostate cancer. Biochem Biophys Acta 1998;1423:F1-F13.

5. Hruban RH, Petersen GM, Goggins M, Tersmette AC, Offerhaus GJA, Falatko F, Yeo CJ, Kern SE. Familial pancreatic cancer. Ann Oncol 1999;10:69-73.

6. Horwitz M. The genetics of familial leukemia. Leukemia 1997;11:1347-1359.

7. Bromen K, Pohlabein H, Jahn I, Ahrens W, Jöckel KH. Aggregation of lung cancer in families: Results from a population-based case-control study in Germany. Am J Epidemiol 2000;152:497-505.

8. Nelson CL, Sellers TA, Rish SS, Potter JD, McGovern PG, Kushi LH. Familial clustering of colorectal, breast, uterine, and ovarian cancers as assessed by family history. Genet Epidemiol 1993;10:235-2444.

9. Vasen HFA, Wijnen JT, Menko FH, Kleibeuker JH, Taal BG, Griffioen G, Nagengast FM, Meijers-Heijboer, EH, Bertario L, Varesco L, Bisgaard M-L, Mohr J, Fodde R and Khan PM. Cancer risk in families with Hereditary Nonpolyposis Colorectal Cancer diagnosed by mutation analysis. Gastroenterology 1996;110:1020-1027.

10. The Breast Cancer Linkage Consortium. Cancer risks in BRCA2 mutation carriers. J Natl Cancer Inst 1999;91:1310-1316.

11. Schildkraut JM, Risch N, Thompson WD. Evaluating genetic association among ovarian, breast and endometrial cancer: Evidence for a breast/ovarian cancer syndrome. Am J Hum Genet 1989;45:521-529.

12. Sutcliffe S, Pharoah PDP, Eason DF, Ponder BAJ, UKCCCR Familial Ovarian Cancer Study Group. Ovarian and breast cancer risks to women in families with two or more cases of ovarian cancer. Int J Cancer 2000;87:110-117.

13. Ford D, Easton DF, Bishop T, Narod SA, Goldgar DE, Breast Cancer Linkage Consortium. Risk of cancer in BRCA1-mutation carriers. Lancet 1994;343:692-695.

14. Dong C, Hemminki K. Modification of cancer risks in offspring by sibling and parental cancers from 2,112,616 nuclear families. Int J Cancer 2001;92:144-150.

15. Anton-Culver H, Ziogas A, Finkelstein D, Griffin C, Hanson J, Isaacs C, Kasten-Sportes C, Mineau G, Nadkarni P, Potter JD, Rimer B, Schildkraut J, Strong L, Weber B, Winn D, Hiatt R, Nayfield S. The Cancer Genetics Network: Recruitment results and pilot studies. Community Genet 2003;6:171-177.

16. De Vita VT, Hellman S, Rosenberg SA, editors. CANCER: Principles and Practice. 5th ed: Lippincott-Raven;1997.

17. Hudson J, Laird N, Betensky R. Multivariate logistic regression for familial aggregation of two disorders: I. Development of models and methods. Am J Epidemiol 2001;153:500-505.

18. Cox DR, Wermuth N. A note on the quadratic exponential binary distribution. Biometrika 1994;81:403-408.

19. Betensky RA, Whittemore AS. An analysis of correlated multivariate binary data: Application to familial cancers of the ovary and breast. Appl Statist 1996;45:411-429.

20. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;83:13-22.

21. Tomizawa Y, Adachi J, Johno T, Yamaguchi N, Saito R, Yokota J. Identification and characterization of families with aggregation of lung cancer. Jpn J Clin Oncol 1998;28:192-195.

22. Bevan S, Catovsky D, Marossy A, Matutes E, Popat S, Antonovic P, Bell A, Berrebi A, Gaminara EJ, Quabeck K, Ribeiro I, Mauro FR, Stark P, Sykes H, van Dongen J, Wimperis J, Wright S, Yuille MR, Houlston RS. Linkage analysis for ATM in familial B cell chronic lymphocytic leukaemia. Leukemia 1999;82:775-781.

23. Bevan S, Catovsky D, Matutes E, Antunovic P, Auger MJ, Ben-Bassat I, Bell A, Berrebi A, Gaminara EJ, Junior ME, Mauro FR, Quabeck K, Rassam SMB, Reid C, Ribeiro I, Stark P, van Dongen JJM, Wimperis J, Wright S, Marossy A, Yuille MR, Hourlston RS. Linkage analysis for major histocompatibility complex-related genetic susceptibility in familial chronic lymphocytic leukemia. Blood 2000;96:3982-3984.

24. Colditz GA, Martin P, Stamfer MJ, Willett WC, Sampson L, Ronser B, Hennekens CH, Speizer FE. Validation of questionnaire information on risk factors and disease outcomes in a prospective cohort study of women. Am J Epidemiol 1986;123:894-900.

25. Mitchell RJ, Brewster D, Campbell H, Porteous MEM, Wyllie AH, Bird CC, Dunlop MG. Accuracy of reporting family history of colorectal cancer. Gut 2004;53:291-295.

26. Rothman KJ, Greenland S. Modern Epidemiology. 2nd ed.: Lippincott Williams and Wilkins; 1998.

27. Gloeckler Ries LA, Koasry CL, Hankey BF, Miller BA, Harras A, Edwards BK, editors. SEER Cancer Statistics Review, 1973-1994. Bethesda (MD): National Cancer Institute; 1997.

28. Cancer Genetics Network Statistical Coordinating Center
http://hedwig.mgh.harvard.edu/cgnpub/

29. National Cancer Institute Cancer Genetics Network Site
http://epi.grants.cancer.gov/CGN/