Published in final edited form as:

Sci Transl Med. 2012 May 23; 4(135): 135le3–135lr3. doi:10.1126/scitranslmed.3004162.

Predictive Capacity of Genome Sequencing

Colin B. Begg and Malcolm C. Pike

Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York

In their interesting and provocative study Roberts et al.¹ address the potential of future genetic epidemiologic investigations using modern tools such as genome sequencing to substantially improve our ability to identify in advance people who will ultimately succumb to a variety of common diseases, reviving a generation old debate about the merits of research on epidemiology and prevention.² Their analyses, involving data from large twin registries, paints a pessimistic picture of the likely fruitfulness of future research in this field. However, the authors have misinterpreted the twin data and have seriously underestimated the potential rewards of continued research into the causes of these diseases.

An ultimate goal of epidemiological research is to accurately determine the disease risk for each individual in the population. The immediate goal addressed by Roberts et al. is to pose a question that is primarily relevant to science policy: how much inherent risk variation from person to person exists in the population, and is thus subject to discovery by future research efforts? The authors are correct in stating that the distribution of risks in the population is the critical measure of relevance. However, the complexity of their methods obscures the essential feature of this risk distribution that can be derived from MZ twin data. It is easily shown that the fundamental measure of risk variation, the coefficient of variation of the population risk distribution, is directly related to the mean risk among cases who contract the disease divided by the population mean risk.³ Identical twins offer the opportunity for a natural experiment to estimate this quantity, since each individual is a replicate of his/her MZ twin. As a result the coefficient of risk variation can be estimated very simply using $(4an/(2a+b)^2 - 1)^{1/2}$, where a is the number of disease concordant pairs, b is the number of discordant pairs, and n is the total number of twin pairs (for proof see Appendix). In the attached table it is shown that the estimates of these coefficients of variation are substantial for most of the 24 diseases studied.

To interpret these results we focus on breast cancer as an example. Breast cancer is an unusual model in that we can validate the results from the twin registries by using an analogous strategy with much more statistical precision. Rather than requiring an identical twin we can match every woman with herself by considering each breast to be identically susceptible. The occurrence of independent second (contralateral) breast cancers are recorded routinely in cancer registries, providing us with large population-based datasets that can be used to measure the inherent aggregation of breast cancer risk within individuals, after adjusting for the fact that only one breast is at risk for the second primary while both breasts are at risk for the first primary. In this context, the ratio of the mean risk in cases to the mean population risk is known as the standardized incidence ratio (SIR), and is traditionally adjusted for age due to the strong dependence of risk on age. Using data from the US SEER (Surveillance Epidemiology and End Results) cancer registries it has been shown that the derived estimate of the SIR for breast cancer is 3.9, leading to an estimate for the coefficient of risk variation of approximately 1.7.⁴ The SIR estimate tells us that, on average, a typical woman diagnosed with breast cancer harbors at the outset a risk approximately 3.9 times greater than a woman with average risk. This result is very similar to the result from the twin registries shown in the table (SIR=4.1), but is based on a vastly

greater sample size. Note that exposures to environmental risks are matched in this method, whereas for twin data the degree of sharing of these exposures is uncertain.

A coefficient of variation of 1.7 ($SIR=3.9$) may not appear large at first glance, but it points to a very large range of risks in the population. While knowledge of the coefficient of variation does not allow us to map out the exact shape of the distribution of risks substantial risk variation ensures that the risk distribution in the population is inevitably highly skewed, with the bulk of the population having considerably lower than average risks, and a relatively small subset having greatly increased risks. For example, if we assume that the risks have a lognormal distribution, a device used by other investigators for similar purposes,⁵ we can display the full population risk distribution. In the figure these are displayed for a variety of SIRs. For each curve the value of 1 on the horizontal axis represents the mean population risk.

These results offer the promise that further research will ultimately allow us to identify the relatively small portion of the population at greatly elevated risk, i.e. those who can benefit from more intensive screening or other prevention strategies. An appropriate tool for characterizing these public health implications is the Lorenz curve, where increasing proportions of the population ranked on the basis of risk are plotted against the proportion of disease occurrences that will happen in the designated high risk segment of the population.^{3,6} Using the lognormal approximation we can infer, for example, that 69% of all breast cancers will occur in the 25% of the population that possess the highest breast cancer risks, and that only 3% of breast cancers will occur in the 25% of the population with the lowest risks. This potential predictability compares very favorably with models based on current known risk factors for which only approximately 40% of all breast cancers are predicted to occur in the 25% of the population with the highest predicted risk.⁷ A further examination of the table shows that for most disease types the twin data suggest even greater risk concentrations than for breast cancer. These results provide promise for the cost-effectiveness of future, targeted disease prevention strategies, and offer a much more optimistic scenario than the one presented by Roberts et al.

References

1. Roberts NJ, Vogelstein JT, Parmigiani G, Kinzler KW, Vogelstein B, Velculescu VE. The predictive capacity of personal genome sequencing. *Sci Transl Med*. Apr 2.2012 10.1126/scitranslmed.3003380
2. Bailar JC 3rd, Smith EM. Progress against cancer? *N Engl J Med*. 1986; 314:1226–1232. [PubMed: 3702918]
3. Begg CB, Satagopan JM, Berwick M. A new strategy for evaluating the impact of epidemiologic risk factors for cancer with application to melanoma. *J Am Stat Assoc*. 1998; 93:415–426.
4. Begg CB. The search for cancer risk factors: when can we stop looking? *Am J Public Health*. 2001; 91:360–364. [PubMed: 11236398]
5. Pharoah PD, Antoniou AC, Easton DF, Ponder BA. Polygenes, risk prediction, and targeted prevention of breast cancer. *N Engl J Med*. 2008; 358:2796–2803. [PubMed: 18579814]
6. Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, Hsieh LJ, Begg CB. Variations in lung cancer risk among smokers. *J Natl Cancer Inst*. 2003; 95:470–478. [PubMed: 12644540]
7. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, Thun MJ, Cox DG, Hankinson SE, Kraft P, Rosner B, Berg CD, Brinton LA, Lissowska J, Sherman ME, Chlebowski R, Kooperberg C, Jackson RD, Buckman DW, Hui P, Pfeiffer R, Jacobs KB, Thomas GD, Hoover RN, Gail MH, Chanock SJ, Hunter DJ. Performance of common genetic variants in breast-cancer risk models. *N Engl J Med*. 2010; 362:986–993. Erratum in: *N. Engl. J. Med*. 363, 2272 (2010). [PubMed: 20237344]

Appendix

Let the population consist of n twin pairs, and let each twin pair possess a genotype with a distinct disease risk, denoted r_i for the i^{th} twin pair. We seek to estimate the coefficient of variation of the distribution of these risks from the two observed frequencies at our disposal: a , the number of disease concordant pairs, and b , the number of disease discordant pairs. Let τ be the coefficient of variation, i.e. $\tau^2 = (\sum r_i^2/n)/\mu^2 - 1$, where μ is the population mean risk. Since the probability that the i^{th} pair will be disease concordant is r_i^2 , and the probability that the pair is discordant is $2r_i(1-r_i)$ it follows that $E(a) = \sum r_i^2$ and $E(b) = \sum 2r_i(1-r_i)$. Thus “ a ” serves as a moment estimator of $\sum r_i^2$ and $(a+b/2)/n$ serves as a moment estimator of μ . Consequently τ can be estimated using $[(4an/(2a+b)^2 - 1)]^{1/2}$.

Figure. Standardized Lognormal Risk Distributions¹

¹. Each curve represents the population distribution of risks under the assumption that the distribution is lognormal, where the value of 1 on the horizontal axis represents the population mean risk. Thus, as the SIR increases, the distribution becomes more and more spread out, with the large preponderance of the population experiencing risks much lower than the average population risk, and a small proportion having increasingly larger risks.

Table

Standardized Incidence Ratios and Coefficients of Risk Variation from Twin Data¹

Disease	# MZ Twins	# Disease Concordant	# Disease Discordant	Standardized Incidence Ratio ^{2,4}	Coefficient of Risk Variation ^{3,4}
Bladder Cancer	15668	5	189	7.9	2.6
Breast Cancer	8437	42	505	4.1	1.8
Colorectal Cancer	15668	30	416	8.3	2.7
Leukemia	15668	2	103	10.9	3.2
Lung Cancer	15668	18	296	10.2	3.0
Ovarian Cancer	8437	3	125	5.9	2.2
Pancreatic Cancer	15668	3	123	11.3	3.2
Prostate Cancer	7231	40	299	8.1	2.7
Stomach Cancer	15668	11	223	11.5	3.2
Thyroid Autoimm.	284	7	17	8.3	2.7
Type 1 Diabetes	4307	3	20	76	8.7
Gallstone Disease	11073	112	956	3.6	1.6
Type 2 Diabetes	4307	29	113	17.1	4.0
Alzheimers*	398	2	8	22	4.6
Dementia	398	3	16	9.9	3.0
Parkinsons*	3477	7	60	17.8	4.1
Chron. Fatigue (F)	1803	133	526	1.5	0.7
Chron. Fatigue (M)	1426	48	266	2.1	1.0
GERD (F)	1260	63	284	1.9	0.9
GERD (M)	918	32	185	1.9	0.9
Irritable Bowel	1252	14	97	4.5	1.9
CHD Death (F)	2004	97	424	2.0	1.0
CHD Death (M)	1640	153	451	1.8	0.9
Stroke Death	3852	35	316	3.6	1.6
General Dystocia	928	40	173	2.3	1.1
Pelvic Prolapse	3376	34	157	9.1	2.8

Disease	# MZ Twins	# Disease Concordant	# Disease Discordant	Standardized Incidence Ratio ^{2,4}	Coefficient of Risk Variation ^{3,4}
Stress Urinary Inc.	3376	13	87	13.7	3.6

¹Data from Roberts et al.¹

²The mean risk in cases with disease divided by the mean population risk.

³The standard deviation of risks in the population divided by the mean population risk.

⁴The precision of the estimates of SIR and CV are strongly dependent on the numbers of disease concordant and discordant twin pairs.