William D. Dupont, Ph.D.

Supporting Statement

A. Statistical Applications and Data Collection

Dr. Dupont has made many contributions to the application of statistical science to the design, conduct, and analysis of observational medical studies and clinical trials. His curriculum vitae lists 131 peer-reviewed publications that have been cited 7,730 times in the literature; his *h*-index (Hirsch index) is 43 (*ISI Web of Science*, December 2008). He is best known for his contribution to the epidemiology of breast disease. In this work he has collaborated with David L. Page, MD, a histopathologist, and other experts in molecular biology and genetics. He and his colleagues have carefully designed, conducted, and analyzed a retrospective cohort study of women at three hospitals in Nashville, Tennessee who have undergone breast biopsies revealing benign breast lesions. Today this cohort includes 17,000 women who have been followed for an average of 17 years. In the 1970s, women who had undergone benign breast biopsy were known to have an elevated risk of breast cancer and were diagnosed with what was then called fibrocystic disease. Dupont & Page (1985) showed that most of these women were not at elevated breast cancer risk and that a minority who had benign proliferative disease had a two-fold elevation in subsequent breast cancer risk. A further minority had atypical hyperplasia, which is associated with a four-fold elevation in breast cancer risk. (Today 10% of women undergoing mammographically directed biopsies have atypical hyperplasia.) This work has redefined benign breast disease, the management of women with benign breast lesions, and the line between atypical hyperplasia and low-grade carcinoma in situ. Their research has been replicated by epidemiologists at Harvard University and the Mayo Clinic and is today accepted by the College of American Pathologists. Their original paper (Dupont & Page 1985) has been cited over 950 times in the literature.

Dr. Dupont has been the Principal Investigator on five R01 grants from the National Cancer Institute. His current grant on the epidemiology of molecular risk factors for breast cancer is in its 17th year of funding. Today his research is focused on the breast cancer risks associated with the interactions between proliferative breast disease and haplotypes of the TGFβ and EGFR gene families, as well as with genes involved in the synthesis, function, and metabolism of estrogen. This work is being done in collaboration with Jeffry R. Smith, MD, a molecular geneticist whose expertise includes the genotyping of archival formalin-fixed paraffin-embedded tissue.

Dr. Dupont's interest in breast cancer has also led to a meta-analysis of the effects of estrogen replacement therapy (ERT) on breast cancer risk (Dupont & Page 1991). This paper found evidence that low-dose conjugated estrogen therapy of a few years duration used to wean menopausal women from their endogenous estrogens did not appreciably

increase breast cancer risk.[1] His research team also found that short term ERT for the relief of menopausal symptoms had, at most, modest effects on breast cancer risk in women with a history of proliferative breast disease or atypical hyperplasia (Dupont et al. 1999). Dr. Dupont has also contributed to the debate on the optimal age to initiate mammographic screening in women (Dupont WD: Evidence of efficacy of mammographic screening for women in their forties. *Cancer*, 1994; 74:1204-6).

B. Teaching and Dissemination of Statistical Knowledge

Dr. Dupont has taught biostatistics to students in the Masters of Public Health at Vanderbilt since the inception of this program in 1997. This program is intended for clinical fellows who seek an academic career in population-based medical research. Its goal is to provide an introduction to modern methods in biostatistics and epidemiology and to provide students with a mentored opportunity to do a research study in their own area of clinical expertise. Dr. Dupont teaches the more advanced of two graduate courses in biostatistics to these students. In 1997 all of the text books that covered the intermediate level biostatistics of this course assumed that the reader was a student of statistics with a considerably greater background in mathematics and statistics than was the case for most of his clinician students. Dr. Dupont developed a course using the Stata software package that avoided much of the mathematics underlying modern multiple regression methods. The course focuses on understanding the models and assumptions underlying these methods, how to perform residual analyses to evaluate model fit, how to select the most appropriate model for the problem at hand, and how to explain the results of these analyses to a clinical audience. In 2002, Cambridge University Press published Dr. Dupont's text, which is based on this course (Dupont 2002). The topics covered in this text are linear regression, logistic regression, Poisson regression, survival analysis, and analysis of variance. Each topic is covered in two chapters: one introduces the topic with simple univariate examples and the other covers more complex multivariable models. Modern graphical methods are emphasized. The text makes extensive use of a number of real data sets that are posted online at http://biostat.mc.vanderbilt.edu/dupontwd/wddtext together with a table of contents and other information. The first edition went through two printings. The second edition, which has been extensively revised and expanded, will be published in February of 2009.

Dr. Dupont has also helped clinicians understand the relationship between relative and absolute risk. For rare events, this relationship is easily explained. However, when the baseline risk becomes appreciable, this relationship becomes more complex. Dr. Dupont has developed graphical methods to help clinicians obtain an intuitive understanding of this relationship and, in collaboration with Dale Plummer, wrote a program to estimate

[1] Although there is considerable controversy about the effect of hormonal replacement therapy (HRT) on breast cancer risk, this finding is consistent with subsequent studies. For example, the Collaborative Group on Hormonal Factors in Breast Cancer (*Lancet* 1997;350:1047-59) found that among current users, breast cancer risk increased by a factor of 1.023 for each year of HRT, and that there was no significant excess breast cancer risk 5 or more years after the cessation of HRT. The Women's Health Initiative, which studied postmenopausal women who for the most part were free of menopausal symptoms, did not directly address the question of whether breast cancer risk is increased by HRT used to wean women from their endogenous estrogens.

absolute risk given a relative risk, age-specific baseline absolute risk, and age-specific competing mortal hazard from other causes (Dupont & Plummer. *Cancer*, 1996; 77:2193-9, Dupont *Stat Med*, 1989, http://biostat.mc.vanderbilt.edu/RelativeToAbsoluteRisks).

C. Statistical Research

Dr. Dupont's doctoral thesis concerned the estimation of animal abundance using catch-effort data. His approach assumes a competing-risk model of adult deaths and captures, avoids making any assumptions about birth rates or juvenile mortality rates, and allows the user to incorporate an arbitrary number of time-dependent covariates into the natural and catch hazard functions (Dupont 1983a). He used this method to estimate the size of the halibut fishery in the North Pacific Ocean. It has also been used by Novak et al. (*J Wildlife Management*, 1991; 55:31-8) to estimate a white-tailed deer population.

After coming to Vanderbilt Medical School his interests switched to methods with medical applications. He has had a career-long interest in sample size estimation and power calculations. He has published a method for power calculations for matched case-control studies (Dupont 1988) and for linear regression (Dupont & Plummer 1998). He has also designed and supervised the writing of the *PS* program for power and sample size calculations. This program covers most of the more commonly used methods for power calculations in medical statistics. These include methods for survival analysis, paired and independent *t*-tests, linear regression, matched and independent tests with dichotomous outcomes, and the Mantel-Haenszel test. Its interactive graphical user interface was designed to facilitate ease of use and permits the creation of publication-quality graphics. The most recent release of this software automatically generates a description of the each power or sample size calculation that may be cut and pasted into grant applications or other documents. This program, which is freely available on the web, has been widely used throughout the world (see http://biostat.mc.vanderbilt.edu/PowerSampleSize). At the beginning of 2009, it was the first hit in response to a Google search for "power and sample size." His review paper (Dupont & Plummer 1990) that describes the first version of this program had been cited 368 times in the literature as of the beginning of 2009.

Dr. Dupont has had a long-term interest in the foundations of statistical inference. An advocate of the likelihood approach to inference, he has written papers that have investigated the adequacy of the *P* value as a consistent measure of strength of evidence. These include Dupont (1983b) that explores the relationship between sequential stopping rules, sequentially adjusted *P* values and strength of evidence, and Dupont (1986) that explores the sensitivity of *P* values from Fisher's exact test to minor perturbations in 2x2 contingency tables. Dr. Dupont has also contributed to the debate on the validity of the adaptive clinical trial design used to evaluate ECMO therapy in premature infants (Dupont WD. *Stat Science*, 1991; 6:69-71).

Dr. Dupont has contributed to the development of statistical graphics with his paper on the density distribution sunflower plot (Dupont & Plummer 2005). This graphic is intended for bivariate data that are too dense to be effectively displayed by scatter plots due to overstrike problems. It combines ideas from Carr et al. (*J Am Stat Assoc* 1987)

with the original sunflower plot of Cleveland & McGill (*J Am Stat Assoc* 1984) to produce a graph that gives both an intuitive sense of the density distribution of the data while also providing detailed information about the number and locations of observations. It generates a grid of small regular hexagonal bins. Data in low-density regions are displayed as in a regular scatter plot, data in medium-density regions are depicted by light sunflowers with one observation per petal, and high-density regions are displayed by dark sunflowers with each petal representing *k* observations. This permits the viewer to determine the exact location of observations in low-density regions, the number of observations in bins in medium-density regions and the number of subjects to within *k* observations in bins in high-density regions. The colors of light and dark sunflowers can be chosen in such a way that the color saturation increases with increasing density of observations. This provides an overall sense of the density distribution of the data. A version of this program has been implemented by Stata Corp. in their statistical software package. (See also http://www.stata.com/meeting/3nasug/sunflower.pdf .)

Dr. Dupont is applying modern statistical methods to the exploratory analysis of the genomic and histologic data that he and his colleagues are collecting in their retrospective cohort study of women with benign breast disease. The large number of SNPs and haplotypes under consideration in this study creates profound multiple comparisons problems. These are being addressed through the use of permutation tests to adjust *P* values and bootstrap analyses to perform appropriate shrinkage adjustments for odds ratio estimates. They will also be using other databases as test sets to validate their findings. Of particular importance is a similar cohort at the Mayo Clinic that is being studied by Dr. Lynn C. Hartmann and colleagues. These research teams at Vanderbilt and the Mayo Clinic will use each others' data as test sets for their exploratory analyses. In his initial analyses Dr. Dupont will follow the supervised principal components (SPC) approach of Bair et al. (*J Am Stat Assoc* 2006;101(473):119-137). This approach will be adapted for use with conditional logistic regression using categorical covariates. False discovery rate methods will be used to avoid serious Type II errors (Storey et al. *PNAS* 2003;100(16):9440-5).

D. Activities in Support of the Statistical Profession

Dr. Dupont has served as a referee for 28 statistical, clinical and other journals. He is an associate editor of the *Stata Journal* and has served as an ad hoc member of numerous NIH Study Sections and other peer review committees. Throughout his 31-year career at Vanderbilt he has worked to foster the application of sound statistical methods to medical research and to enhance the professional environment of his statistical colleagues. From 1989 until 2003 he was the Director of the Division of Biostatistics at Vanderbilt University. During this time, much of the senior leadership at Vanderbilt were unconvinced of the need to provide institutional support for biostatistics. Dr. Dupont worked to support his Division during these years and, with others, lobbied for institutional support for biostatistics at Vanderbilt. This eventually led to the creation of the Department of Biostatistics at Vanderbilt which, under the leadership of Dr. Frank E. Harrell, has expanded to include 28 full-time faculty and 21 staff biostatisticians.

Selected References

1. Dupont WD: A stochastic catch-effort method for estimating animal abundance. *Biometrics,* 1983a; 39:1021-33.
2. Dupont WD: Sequential stopping rules and sequentially adjusted P values: Does one require the other? *Controlled Clin Trials,* 1983b; 4:3-10. Rejoinder: 4:27-33.
3. Dupont WD and Page DL: Risk factors for breast cancer in women with proliferative breast disease. *N Engl J Med*, 1985; 312:146-51.
4. Dupont WD: Sensitivity of Fisher's exact test to minor perturbations in 2 x 2 contingency tables. *Stat Med*, 1986; 5:629-35.
5. Dupont WD: Power calculations for matched case-control studies. *Biometrics*, 1988; 44:1157-68.
6. Dupont WD: Converting relative risks to absolute risks: A graphical approach. *Stat Med*, 1989; 8:641-51.
7. Dupont WD and Plummer WD: Power and sample size calculations: A review and computer program. *Controlled Clin Trials*, 1990; 11:116-28.
8. Dupont WD and Page DL: Menopausal estrogen replacement therapy and breast cancer. *Arch Intern Med*, 1991; 151:67-72.
9. Dupont WD and Plummer WD: Power and sample size calculations for studies involving linear regression. *Controlled Clin Trials*, 1998; 19: 589-601.
10. Dupont WD, Page DL, Parl FF, Plummer WD, Schuyler PA, Kasami M, and Jensen RA: Estrogen replacement therapy in women with a history of proliferative breast disease. *Cancer*, 1999; 85: 1277-83.
11. Dupont WD, Plummer WD: Using density distribution sunflower plots to explore bivariate relationships in dense data. *Stata Journal*, 2005; 5(3):371 - 84.
12. Dupont WD: *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data.* Cambridge U.K.: Cambridge University Press, 2002. 2nd Edition 2009.