Why worry about awareness in choice problems? Econometric analysis of screening for cervical cancer*

Rochelle Belkar[1], Denzil G. Fiebig[1,2], Marion Haas[2] and Rosalie Viney[2]

[1]School of Economics, University of New South Wales
[2]CHERE, University of Technology, Sydney

Abstract: Cervical cancer is one of the most preventable and curable forms of cancer. Since 1991 there has been a concerted effort in Australia to recommend and encourage women to have Pap smears every two years. Part of the success of this National Cervical Screening Program can be gauged by exploring the determinants of screening for cervical cancer among high-risk women and by addressing the specific question of whether screening is associated with socio-economic status. Accessibility to health services remains a core goal in health policy in Australia but evidence on whether the goal is being met is limited. Using unit record data from the 1995 National Health Survey, an econometric model is developed for whether women have ever screened or not. A proportion of women in the sample contend that they have never heard of a Pap test. The analysis characterizes this group of women and accounts for their presence in our modelling.

Corresponding Author:
Professor Denzil G Fiebig
School of Economics
UNSW
Sydney, 2006
Australia
Ph +61-2-9385 3958 Fax +61-2-9313 7691
Email d.fiebig@unsw.edu.au

Version: 2 February, 2004

1. Introduction

Most research in consumer decision-making concentrates on the ultimate choice of whether to purchase or not when in fact this final decision is likely to be the result of a hierarchical system of decision states. In the first instance, consumers need to be aware of the existence of the product or service. Then they will typically acquire product information leading to some basic level of interest or disinterest. Only then is it possible for the product to be in the consumer's choice set at which time there is a further set of decisions: conditional on wanting to purchase, can they afford it and if so when and if so what brand or type.

It is easy to see the applicability of such a structure in the health context. New tests and products are continually being developed and knowledge about these will only gradually filter through to doctors and consumers. Even in the case of cervical screening, which is a relatively mature and well-known service, recent national surveys in Australia indicated that a substantial proportion (4% in 1995) of women had never heard of a Pap test. For these women, they have never had a test not because of any conscious choice but because they simply were not aware of its existence.

From an econometric perspective the issue of whether a woman is aware of Pap tests presents a potential selection problem. What is of interest is a conditional probability of choosing to test given awareness. It is not appropriate to eliminate unaware women from the sample and analyse the choice behaviour of the subset of women who are aware of the Pap test unless the awareness and screening choices are independent. This would give the wrong conditional probability.

There is also a problem if the entire sample is used with no differentiation made between non-screeners who are there by choice and those that are non-screeners because of lack of awareness. This would be the case if, as often happens, the awareness question was not asked and respondents were only asked whether they screened or not. Here the problem is that the hypothesized choice model is only appropriate for a subset of the sample. For those who are not aware, they will be non-screeners irrespective of the values taken by the explanatory variables.

Using unit record data from the 1995 National Health Survey collected by the Australian Bureau of Statistics (ABS), econometric models are developed for whether women have ever screened for cervical cancer or not. Cervical cancer is one of the most preventable and curable forms of cancer. A cervical cancer may take 10 or more years to develop, during which time the cells of the lining of the cervix may show pre-cancerous changes which can be detected by a Pap smear (also called Pap test), allowing for early treatment with an excellent chance of a full recovery.

Since 1991 there has been a concerted effort in Australia to recommend and encourage women to have Pap smears every two years. The National Cervical Screening Program (NCSP) aims to reduce morbidity and deaths from cervical cancer through an organised approach to screening, incorporating state based registries, information for general practitioners and women and quality assurance programs relating to all aspects of cervical cancer screening. The NCSP encourages women in the target population to have regular Pap smears and is jointly funded by Commonwealth and State/Territory governments.

Deaths from cervical cancer fell by 40% between 1986 and 1998. Since the NCSP was first introduced in 1991, cervical cancer has fallen from the eighth to the fourteenth most common cancer among Australian women and participation in screening, defined as those women who regularly screen, is now close to 70%. In addition, performance standards and guidelines have been developed for laboratories reporting cervical cytology and histopathology, management and follow-up of women with screen-detected abnormalities and colposcopies. Since 1999, all States and Territories have in place cervical cytology registers (Pap Test Registers or PTR).

In assessing the success of the NCSP it is relevant to explore the determinants of screening for cervical cancer among women and the specific question of whether screening is associated with socio-economic status. Accessibility to health services remains a core goal in health policy in Australia but evidence on whether the goal is being met is limited.

While there is considerable interest in frequency of screening and the NCSP has emphasised screening every two years, ensuring wide coverage is also a major

priority. Approximately 30% of women still do not have regular Pap smears. According to the NCSP, three out of four women who develop cervical cancer each year have never had a Pap smear or not had one within the recommended two yearly interval (see http://www.cervicalscreen.health.gov.au/facts/index.html). Thus explaining initial screening choice is an important aspect of screening behaviour.

While the number of women who are unaware of pap tests is not large, they represent a large proportion of those women who have never screened, and the most recent data suggest that a substantial minority remain unaware. As such we also investigate the sensitivity of the econometric analysis and the resultant inferences to the treatment of those women in the sample who contend that they have never heard of a Pap test. Of independent interest is the characterisation of this group of women. Again, given coverage is a key objective of the NCSP, these are an important group of women to study.

2. Demand for screening

The theoretical framework for the empirical work follows from Kenkel's (1990, 1991, 1994, 2000) analysis of the demand for preventive care. Consumption of health care is regarded as an investment in health human capital (Grossman, 1972). However, treatment services provide both immediate benefits in terms of symptom relief and a potential increase in the stock of health capital. Preventive care can be regarded more purely as an investment in an individual's stock of health capital. It is more appropriately modelled under conditions of uncertainty.

In defining preventive medical care, Kenkel (2000) distinguishes between primary prevention, which reduces the probability of illness, and secondary prevention, which reduces the consequences of illness through early detection, without necessarily affecting the occurrence of disease. Screening services are secondary prevention. For example, cervical screening is aimed at detection and treatment of pre-cancerous lesions, avoiding the development of invasive cancer. Secondary prevention has also been seen as "self-insurance" (Erlich and Becker, 1972), pointing to a possible substitution between preventive care and curative care.

Broadly within a health human capital theoretical framework, Kenkel (1990, 1994, 2000) identifies and tests empirically a range of factors that are likely to determine demand for secondary prevention services such as cervical screening. Increasing age is predicted to increase the demand for many preventive services because of increased risk of disease (depreciation of health capital), but this is offset by the potential impact of a perceived shorter payoff period for any investment. Kenkel (1994) found that the demand for preventive care (measured as use of breast and/or cervical cancer screening) decreased with age. Higher incomes are also expected to be associated with higher demand for preventive services, not just because of an income effect, but also because higher incomes are indicative of greater investment in human capital, and therefore higher costs associated with depreciation of human capital through ill health.

The relationship between insurance coverage and demand for preventive care is complicated by issues of whether the insurance is for curative care or also covers preventive care and the extent to which preventive care is self insurance. That is, there will be cross price effects if insurance reduces the cost of both preventive and curative care, but curative care and secondary prevention may be complements rather than substitutes. Further, insurance coverage may be endogenous. Empirically, insurance is generally found to increase demand for preventive care or preventive behaviour (Kenkel, 1994).

Of particular interest to the current study is the impact of health information, knowledge and schooling on demand for preventive care. It has generally been found that higher levels of schooling are associated with higher demand for medical care, particularly preventive care and with increased preventive health behaviours (Kenkel, 1991, 1994; Mullahy, 1999). Higher levels of education may increase the efficiency of household production of health, but the effect of this on demand for preventive care is ambiguous.

What decisions an individual consumer makes about preventive care is likely to be influenced by the type and amount of information she receives and is able to use. This, in turn, may be related to higher levels of education. For example, education may reduce the costs of acquiring health information and may increase an individual's ability to understand and act on information about preventive care. Kenkel (1990) has

shown that while more informed consumers are more likely to visit the doctor, information seems to have little impact on how much care is consumed. His research also showed that women who are better informed are more likely to demand preventive care. Education may be related to screening via other mechanisms. Kenkel (2000) hypothesises that schooling may influence demand for preventive care via an improvement in allocative efficiency, that is better knowledge about the relationship between health inputs and health will lead to increased allocation of resources to health inputs. An alternative explanation of the impact of education is that both higher levels of schooling and higher demand for preventive care are indicative of a lower rate of time preference.

The current study considers the relationship between awareness of cervical screening and demand for cervical screening. Awareness is a necessary but not sufficient condition for screening. Awareness will be related to health knowledge, to factors affecting the acquisition of health information and to education, but is also likely to be influenced by other factors such as culture and language.

3. Modelling awareness and choice

In order to translate the demand for screening into an econometric model consider two latent variables representing the propensities of the ith woman to be aware of Pap tests and to choose to have a Pap test. Denote the latent variable representing choice by y_1* and that representing awareness by y_2* and assume a model specification of the form:

(1) $\quad y_{ji}* = x'_{ji}\beta_j + u_{ji} \, ; j = 1, 2$

where the realization of the latent variable $y_{ji}*$ is defined by $y_{ji} = 1$ if $y_{ji}* > 0$ and $y_{ji} = 0$ otherwise and x_{ji} are vectors of socio-demographic characteristics affecting the choices of the ith woman. Under the assumption that the disturbance term (u_{1i}, u_{2i}) is bivariate normal with:

var (u_{ji}) = 1 and cov (u_{1i}, u_{2i}) = ρ

(1) specifies a bivariate probit model. The sample selection or censored probit variant of this model occurs when both y_1 and y_2 are observed only if $y_2 = 1$. Thus, only for

those women who are aware of Pap tests do we observe their conscious choice of screening or not. Several other variants of a bivariate probit model with partial observability are discussed in Meng and Schmidt (1985).

Given this structure, univariate probit analyses of screening choice will typically produce inconsistent estimates of the model's coefficients. If the available data allow us to separate those women who are aware and this sub-sample is used to estimate the screening model then this "sub-sample" probit analysis will produce inconsistent parameter estimates unless $\rho = 0$. Only in this case will there be no self-selection problem. In some cases there is no information that enables this split of the sample. If the awareness question is not asked and women are simply asked whether they have ever had a Pap test or not then those women who are unaware will be treated as non-screeners. In this case, "full-sample" probit will also produce inconsistent parameter estimates because of the inflation of the number of non-screeners in the sample.

It is important to emphasize here that we are interested in potential outcomes. If unaware women were made aware of Pap tests, would they choose to screen or not? Only in the extreme case that all unaware women would choose not to screen would the "full-sample" probit yield consistent parameter estimates. This concentration on potential outcomes is somewhat controversial. In health economics, it was a key part of vigorous exchanges between proponents of two-part and self-selection models; see Jones (2000) for a summary.

A simple example will provide us with a comparison of these three alternative approaches to the estimation of screening behaviour and of the factors affecting the extent of the inconsistencies. Previous studies of sample selection with a continuous censored variable [see for example Nelson (1984), Zuehlke and Zeman (1991) and Leung and Yu (1996)] have isolated the degree of censoring and dependence between the disturbances as key factors in any comparison of alternative estimators. Suppose there is a single explanatory variable x governing both awareness and screening and that x is a dummy variable taking on a value of unity with probability 0.5. This abstracts from another key factor in these comparisons, namely the degree of correlation between the explanatory variables in the censoring and outcome equations. However, this simplification enables us to more easily isolate some stylised results.

The simple formulation of (1) is given by:

(2) $\quad y_{1i}^* = \beta_{11} + \beta_{12}x_i + u_{1i}$
$\quad\quad y_{2i}^* = \beta_{21} + \beta_{22}x_i + u_{2i}$

Given parameter values $\beta_{11}= 0$, $\beta_{12}= 1$, $\beta_{22}= -1$, there are two parameters that are left to vary. The degree of censoring is governed by β_{21} while the correlation between screening choice and awareness is measured by ρ. Table 1 provides the probability limits of sub-sample and full-sample probit estimates of the parameter of interest (β_{12}) for a range of β_{21} and ρ values. The MLE associated with the censored probit model serve as the base as they are consistent producing a probability limit equal to the true value of β_{12} which has been set to unity. See Appendix 1 for further details.

These results indicate that the inconsistency of both sub-sample and full-sample probit increases with the degree of censoring (as indicated by smaller values of β_{21}). The inconsistency of sub-sample probit increases with the magnitude of the correlation between awareness and screening choice (as measured by ρ). Full-sample probit always underestimates β_{12} and the degree of underestimation varies inversely with ρ. Sub-sample probit typically dominates full-sample probit except when ρ is close to unity.

Table 1: Inconsistency of alternative probit estimators

ρ	$\beta_{21}= 1$		$\beta_{21}= 2$		$\beta_{21}= 3$	
	Sub-sample	**Full-sample**	**Sub-sample**	**Full-sample**	**Sub-sample**	**Full-sample**
-1.0	0.48	-0.41	0.88	0.48	0.98	0.91
-0.8	0.51	-0.39	0.88	0.48	0.98	0.91
-0.6	0.60	-0.35	0.89	0.48	0.98	0.91
-0.4	0.72	-0.30	0.91	0.49	0.99	0.91
-0.2	0.85	-0.25	0.95	0.52	0.99	0.92
0.0	1.00	-0.20	1.00	0.55	1.00	0.92
0.2	1.17	-0.15	1.07	0.59	1.01	0.94
0.4	1.39	-0.10	1.15	0.63	1.03	0.95
0.6	1.71	-0.06	1.27	0.69	1.05	0.97
0.8	2.29	-0.01	1.46	0.77	1.08	0.99
1.0	nd	0.00	nd	1.00	1.08	1.00

Notes: (a) Entries represent probability limits of alternative estimators of $\beta_{12}= 1$.
(b) nd = not defined.

In some circumstances, there can be a trade-off between censoring and dependence. The magnitude of the inconsistency can be relatively large even for moderate or minor censoring if at the same time dependence is strong. Only when both censoring and dependence are minor does sub-sample probit produce reliable estimates.

When the sample split into aware and unaware is unknown, full-sample probit is associated with small inconsistencies only when censoring is minor and dependence is strongly positive. In other cases it may be possible to improve matters by considering alternatives to the full-sample probit. In terms of econometric structure, the problem is similar to that which occurs with a mis-classified binary dependent variable. See discussion by Hausman, Abrevaya and Scott-Morton (1998) and Hsiao and Sun (1999) for alternative estimation procedures in this situation.

4. Data

Data from the 1995 National Health Survey conducted by the Australian Bureau of Statistics (ABS) is used for the current analysis. This is a comprehensive national survey on health-related issues covering approximately 23,000 Australian households. There are a number of exclusions that are performed in order to produce the final data set used in our econometric work. While all participants answered a large common block of questions, the sample is essentially divided into two parts in order to concentrate on different aspects of health. In particular, approximately thirty percent of the NHS sample who responded to a General Health and Wellbeing (SF36) questionnaire were not asked questions relating to private health insurance, educational qualifications obtained since leaving high school, and were not eligible for inclusion in the sub-sample asked questions specifically about women's health issues. A further 138 adult dependents were excluded. These are typically full-time students. After these exclusions the sample consists of 9,232 women. Finally, 263 women were excluded because they did not have valid answers for the questions on Pap smears, leaving a sample of 8,969 that was used in our analyses.

Table 2 provides a brief description of the variables to be used in the analyses. The two endogenous variables are whether or not a woman has ever heard of a Pap test and whether or not they have ever had a Pap test. We expect awareness to be affected by personal characteristics including, age, language proficiency, ethnicity, education

and income. Because individual income is available, the income of the spouse is included as a separate variable rather than aggregating to yield household income. The same characteristics are also likely to impact on screening choice. Additional factors likely to explain screening choice include whether the woman has private health insurance, their family status, their general health status and the presence of risk factors such as whether they are a smoker or not.

Because of item non-response, there is a potential problem with the income variables. The data set contained missing observations on at least one of the income questions for approximately 13% of all observations. At this stage these data have not been excluded but have been identified via two dummy variables (INCMISS and SINCMISS) that were constructed to indicate the presence of missing income and missing spouse income.

Table 2 presents variable means for three relevant sub-samples of the data: those who were unaware of Pap tests and those who were aware of Pap tests divided into screeners and non-screeners. From Table 2 we see that 7,767 women have ever screened representing a screening rate compared to the full sample of 86.6%. While only 4.3% of the total sample were not aware of Pap tests, these 389 women represent 32.4% of the 1,202 women in the sample who have never screened.

There are several key features that distinguish the sub-sample who were unaware of Pap tests from those who were aware. The unaware are much older and are less healthy. They are poorer and, therefore, not surprisingly are less likely to have private health insurance. Because they are much less likely to be Australian born they also tend to have poorer English skills. The most striking difference in terms of country of birth is the relatively large proportion of Asian-born women amongst those who were unaware of Pap tests.

Some, but not all, of the characteristics useful in distinguishing between the unaware and aware women emerge as key discriminators when characterizing non-screeners versus screeners in the aware sub-sample. Those people who screen tend to be better educated, richer and more likely to have private health insurance. However, there is little difference in terms of age, health status, ethnicity and English aptitude.

Table 2: Variable definitions and sub-sample means

Variable	Definition	Unaware	Aware non-screeners	Aware screeners
		n=389	n=813	n=7,767
PAPSMEAR	1 if ever screened	0.00	0.00	1.00
PAPAWARE	1 if aware of pap tests	0.00	1.00	1.00
AGE	Age in years	50.09	44.94	43.19
AGE2024	1 if 20-24 years old	0.16	0.22	0.09
AGE2529	1 if 25-29 years old	0.08	0.09	0.12
AGE3034	1 if 30-34 years old	0.06	0.04	0.13
AGE3539	1 if 35-39 years old	0.05	0.04	0.13
AGE4044	1 if 40-44 years old	0.04	0.04	0.12
AGE4549	1 if 45-49 years old	0.05	0.03	0.11
AGE5054	1 if 50-54 years old	0.03	0.02	0.08
AGE5559	1 if 55-59 years old	0.06	0.03	0.06
AGE6064	1 if 60-64 years old	0.06	0.04	0.05
AGE6569	1 if 65-69 years old	0.10	0.07	0.05
AGE7074	1 if 70-74 years old	0.07	0.08	0.04
AGE7579	1 if 75-79 years old	0.07	0.08	0.02
AGE80PL	1 if 80 years old or more	0.11	0.09	0.01
INC	Personal income $'000	10.28	12.86	15.58
INCMISS	1 if income missing	0.09	0.07	0.10
SINC	Personal income of spouse $'000	6.80	5.97	18.66
SINCMISS	1 if spouse income missing	0.06	0.04	0.08
HLTHEX	1 if self-reported health excellent	0.16	0.18	0.21
HLTHVG	1 if self-reported health very good	0.22	0.36	0.36
HLTHG	1 if self-reported health good	0.37	0.27	0.28
HLTHF	1 if self-reported health fair	0.20	0.14	0.11
HLTHP	1 if self-reported health poor	0.05	0.05	0.03
TERT	1 if tertiary qualifications	0.05	0.10	0.14
DIPLOMA	1 if diploma	0.07	0.06	0.10
TRADE	1 if trade qualification	0.12	0.14	0.18
KIDS	1 if have dependent children	0.18	0.09	0.44
SINGLE	1 if single	0.56	0.69	0.31
HPHI	1 if have private health insurance	0.24	0.33	0.41
SMOKE	1 if current smoker	0.24	0.33	0.48
QUIT	1 if quit smoking	0.14	0.16	0.25
EXENG	1 if able to speak English	0.68	0.94	0.97
ENGLISH	1 if usually speaks English	0.61	0.91	0.95
ENGEXENG	EXENG*ENGLISH	0.58	0.88	0.94
AUBORN	1 if born in Australia	0.48	0.74	0.75
NUBORN	1 if born in New Zealand or UK	0.06	0.10	0.11
SEBORN	1 if born in Southern Europe	0.14	0.04	0.04
WEBORN	1 if born in Western Europe	0.01	0.01	0.02
ASBORN	1 if born in Asia	0.26	0.07	0.04
OTBORN	1 if born in other countries	0.05	0.04	0.04
NSW	1 if reside in NSW	0.18	0.15	0.15
VIC	1 if reside in VIC	0.34	0.27	0.25
SA	1 if reside in SA	0.12	0.18	0.16

Variable	Definition	Unaware	Aware non-screeners	Aware screeners
WA	1 if reside in WA	0.11	0.12	0.10
TAS	1 if reside in TAS	0.04	0.04	0.05
NT	1 if reside in NT	0.07	0.04	0.07
ACT	1 if reside in ACT	0.07	0.10	0.09
BRISBANE	1 if reside in Brisbane	0.04	0.05	0.06
OTQLD	1 if reside in QLD but not Brisbane	0.03	0.04	0.05
METRO	1 if reside in metropolitan area	0.65	0.58	0.53
RURAL	1 if reside in rural area	0.05	0.08	0.07
REMRURAL	1 if reside in remote rural area	0.10	0.11	0.14

For several of the characteristics, the pattern across the three sub-samples is less clear-cut. In terms of family structure, aware screeners are much less likely to be single than either those women who are unaware or those who are aware but have never screened. Screeners are also much more likely to have children and to be smokers than the other two groups. Given identified risk factors associated with cervical cancer these patterns are not surprising.

5. Estimation results

Estimation results for our censored probit model are provided in Table 3. The estimate for ρ indicates strong positive correlation between awareness and choice. The associated Wald-test for ρ suggests a highly significant estimate although Nawata and McAleer (2001) demonstrate, in the context of a standard self-selection model, that this test tends to over-reject the null of no correlation. If instead, the LR test is used, as they recommend, the Chi-squared test statistic is 5.09 with an associated *p*-value of 0.024, which also supports the need to use the censored probit approach. The extent of the biases that would result from ignoring this correlation will be discussed in the next section.

In order to counteract the missing income, the modified zero-order technique has been used. The INCMISS and SINCMISS dummy variables were included as explanatory variables in all equations and when missing, the income variables were coded as zero. While we know of no formal evaluations of the use of the modified zero-order technique in discrete choice models, we suspect that the method is likely to perform

well in our type of situation where there are a number of explanatory variables but only a few have missing observations and where the number of affected cases is substantial. See Greene (2000) and Maddala (1977) for further details of this method in the context of linear regression models. As a sensitivity check, the estimation was redone after deleting those observations where either INCMISS = 1 or SINCMISS = 1. The results were qualitatively the same as those reported.

First, consider the awareness equation where the sign pattern of the estimated coefficients largely conforms to a priori expectations. There is a distinct and precisely estimated age profile. Awareness steadily increases until the 40 to 44 age group then declines. The oldest women in the sample, those over 80, were much less likely to be aware of Pap tests than the reference group (18 to 19 year olds). Both English aptitude and whether English is normally spoken at home have large, distinct and significant impacts on awareness. The interaction effect moderates the impact of these two factors although it is not precisely estimated. The Asian effect identified in the summary statistics remains after controlling for all other factors. Asian women are much less likely to be aware of Pap tests relative to all other ethnic groups. Education does have some impact after controlling for all other factors in that women with tertiary education are somewhat more likely to be aware than women with diploma, trade or no qualifications. After controlling for other factors, personal income does not appear as a significant determinant of awareness but the income of the woman's spouse does have a positive and significant impact. Coefficients associated with spatial effects were typically not precisely estimated with Northern Territory having the largest impact and being the only coefficient to be significant. This indicated that women in the Northern Territory were less aware than women in metropolitan New South Wales that served as the base.

For the screening choice equation one needs to be careful with the interpretation of the coefficients. Screening choice needs to be considered jointly with awareness and hence coefficients cannot immediately be linked to marginal responses for the appropriate conditional or joint probability. In the first instance we simply report on the relative size and significance of the coefficients in the screening choice equation and later augment the discussion with an overview of how these results impact on the joint probabilities.

Table 3: Estimation results: Censored probit

Variables	AWARE Coefficient[a]	AWARE Standard Error	SCREEN Coefficient[a]	SCREEN Standard Error
Constant	0.1198	0.1856	-0.9037***	0.2465
AGE2024	0.1582	0.1370	0.7432***	0.1030
AGE2529	0.5684***	0.1547	1.1934***	0.1130
AGE3034	0.7654***	0.1637	1.4378***	0.1275
AGE3539	0.7529***	0.1662	1.3961***	0.1297
AGE4044	0.9269***	0.1833	1.4358***	0.1297
AGE4549	0.7382***	0.1716	1.5810***	0.1292
AGE5054	0.8111***	0.1867	1.9219***	0.1490
AGE5559	0.4537***	0.1653	1.5782***	0.1379
AGE6064	0.3908**	0.1629	1.4844***	0.1350
AGE6569	0.1547	0.1532	1.0932***	0.1230
AGE7074	0.1236	0.1584	0.8906***	0.1220
AGE7579	-0.1725	0.1619	0.5245***	0.1304
AGE80PL	-0.6013***	0.1574	0.1579	0.1413
EXENG	0.5977***	0.1282	0.1306	0.1742
ENGLISH	0.7334***	0.1905	0.2453	0.2200
ENGEXENG	-0.1286	0.2117	0.3018	0.2268
NUBORN	0.1069	0.1098	-0.1104	0.0721
SEBORN	-0.2678**	0.1301	-0.1399	0.1202
WEBORN	0.1898	0.2253	0.3856**	0.1772
ASBORN	-0.8189***	0.1102	-0.6041***	0.1114
OTBORN	-0.0050	0.1361	0.0280	0.1123
TERT	0.2188*	0.1151	0.1204	0.0752
DIPLOMA	-0.0468	0.1058	0.0806	0.0822
TRADE	0.0425	0.0813	0.1106*	0.0609
INC	0.0036	0.0068	0.0145***	0.0046
INQSQ	0.0000	0.0001	-0.0002**	0.0001
INCMISS	0.0912	0.1196	0.2985***	0.0938
SINC	0.0183***	0.0048	0.0030	0.0059
SINQSQ	-0.0002*	0.0001	0.0000	0.0001
SINCMISS	0.3407**	0.1324	-0.0049	0.1292
VIC	-0.0344	0.0834	-0.0757	0.0660
SA	0.1481	0.0993	-0.0839	0.0735
WA	-0.1137	0.1063	-0.2225***	0.0818
TAS	-0.0163	0.1486	0.1490	0.1148
NT	-0.2325*	0.1246	0.1715	0.1156
ACT	0.0965	0.1242	-0.0786	0.0903
BRISBANE	0.0809	0.1505	0.0487	0.1090
OTQLD	0.1645	0.1654	0.0267	0.1140
RURAL	0.0988	0.1245	-0.0542	0.0857
REMRURAL	-0.0387	0.0919	-0.0142	0.0691
HPHI			-0.0030	0.0457
KIDS			0.6034***	0.0732
SINGLE			-0.3692***	0.0904
SMOKE			0.2305***	0.0577
QUIT			-0.0269	0.0666

Variables	AWARE		SCREEN	
	Coefficient[a]	Standard Error	Coefficient[a]	Standard Error
HLTHEX			0.0821	0.1072
HLTHVG			0.0787	0.1013
HLTHG			0.1755*	0.1012
HLTHF			0.0907	0.1070
ρ	0.6867	0.1763		
Sample size	8969			
Censored obs	389			
Log-likelihood	-3228.7			
Chi-squared (49)[b]	1255.4			

Notes: (a) *** significant at 1% level, ** significant at 5% level, * significant at 10% level
(b) Wald test for null hypothesis that all coefficients in the screening equation (except the intercept) are jointly zero; 1% critical value is 57.34.

There is again a distinct and precisely estimated age profile with estimated age coefficients initially increasing before eventually declining. This is consistent with Kenkel's (1994) findings and with a human capital model of demand for screening. Compared with the awareness equation, the peak occurs later in the 50 to 54 age group. Both English aptitude and whether English is normally spoken at home have positive impacts in the screening choice equation but, in contrast to the awareness results, these coefficients are not individually significant and the interaction effect is positive. While these individual coefficient estimates are not precisely estimated, a joint test of the hypothesis that the three language coefficients are each zero is easily rejected. The associated LR test statistic is 32.55 compared to a Chi-squared critical value of 11.34 at the 1% level. The Asian coefficient is again large and precisely estimated. With education, the base category is associated with no qualifications. All other categories indicate some form of education and all are associated with positive coefficients, consistent with Kenkel's (1990, 1994) findings.

Unlike the awareness equation there is a significant personal income effect but no significant impact of spouse's income. Again, this is consistent with a human capital model of preventive services. Having private health insurance is not a significant determinant. Estimated coefficients for being single and having children are both significant, the former being negative and the latter positive. The estimated coefficient associated with current smokers is positive and precisely estimated but there is no significant difference between non-smokers and those that have quit smoking. General health status does not enter as a significant determinant in the screening choice

equation. As with awareness there seem to be few spatial effects with the only significant impact associated with Western Australia.

In order to gain further insights into the magnitude of some of these effects the results have been translated into impacts on probabilities. The probability of each of the three possible outcomes (unaware, aware non-screener and aware screener) was derived for each woman in the sample. Means of these estimated probabilities were then calculated over the entire sample and over selected groups of women. There are alternative ways to make such comparisons. Loosely speaking, this particular choice allows us to say compare an "average" Australian-born woman with an "average" Asian-born woman. These mean probabilities are provided in Table 4.

While some of the age effects seemed quite large in the two estimated equations, levels of awareness and choice to be screened are quite high and hence these effects often do not translate into large probability differences. Women aged between 30 and 60 are very similar. They have extremely high predicted probabilities of being aware and of having screened. As we move away from this group, predicted screening and awareness probabilities do decline substantially. In the case of younger women they are much more likely to be aware non-screeners than older women who tend to be non-screeners because of lack of awareness.

Women who are Asian born or are non-English speaking or who do not usually speak English are at very high risk of being unaware which then translates into very low predicted screening probabilities. Each of these effects is large individually, and in addition the combined impact of being non-English speaking and of not usually speaking English is substantial. It is these women who are most at risk.

The income comparisons have only been provided for couples, where overall, screening rates are higher than average. High income (defined as above the median income) for either of the couple implies high levels of awareness and screening. For low-income couples, awareness and screening are lower than the full sample averages although the magnitudes of the effects are not nearly as large as the impact of being Asian born or not being a natural English speaker.

Table 4: Mean estimated probabilities for awareness and screening[a]

Group	Number in group	Unaware	Aware non-screeners	Aware screeners
Full sample	8969	0.043	0.091	0.865
18-19 years old	238	0.096	0.493	0.411
20-29 years old	1950	0.046	0.127	0.826
30-39 years old	2091	0.021	0.031	0.948
40-49 years old	1835	0.017	0.031	0.952
50-59 years old	1162	0.032	0.034	0.934
60-69 years old	892	0.072	0.098	0.830
70-79 years old	608	0.092	0.220	0.688
80+ years old	193	0.226	0.366	0.408
Australian born	6590	0.028	0.092	0.880
Asian born	469	0.213	0.115	0.672
English speaking[b]	8233	0.027	0.088	0.885
Able to speak English but usually don't	319	0.129	0.147	0.724
Usually speaks English but non-English speaking	106	0.144	0.202	0.653
Non-English speaking	311	0.357	0.094	0.549
High personal, high spouse income[c]	1693	0.009	0.028	0.963
High personal income, low spouse income	910	0.031	0.039	0.930
Low personal income, high spouse income	1038	0.018	0.036	0.945
Low personal, low spouse income	1214	0.071	0.082	0.847

Notes: (a) Estimated probabilities for each of the three possible outcomes are computed for each woman and then averaged over those in the specified group.
(b) The two binary variables that measure English language skills are used to divide the sample into four groups.
(c) Income comparisons are for couples only where high and low incomes are defined relative to median personal and spouse incomes.

6. A comparison of results

Ultimately, the primary objective is to understand screening behaviour. With the type of data available to us, three alternative estimation procedures have been discussed. Results for the screening equation estimated by censored probit have already been provided in Table 3 while sub-sample and full-sample probit are provided in Appendix 2 as Table A1. Taking each of these three sets of estimates, screening probabilities have been estimated for all women in the sample. In Table 5 averages of

these estimated screening probabilities have been calculated for the full sample and for selected sub-groups.

The anticipated biases associated with sub-sample probit translate into large differences in estimated probabilities relative to those produced by censored probit. The former overstates the probability of screening for the entire sample and for each of the subgroups and the differences are quite substantial for the unaware and aware non-screeners.

**Table 5: Comparison of mean estimated screening probabilities**

Group	Number in group	Censored probit	Sub-sample probit	Full-sample probit
Full sample	8969	0.878	0.898	0.866
Unaware	389	0.657	0.737	0.605
Aware	8580	0.888	0.905	0.878
Aware non-screeners	813	0.681	0.721	0.669
Aware screeners	7767	0.910	0.924	0.900

Note: For each woman in each of the specified groups estimated probabilities of screening are determined using each of the three estimation procedures. Numbers reported are averages of these probabilities.

In order to further illustrate these differences, kernel density plots have been generated for the estimated probabilities of screening for the unaware sub-group. These plots are provided in Figure 1. The difference between the censored probit and sub-sample probit plots is marked with the latter predicting a much larger proportion of women to be highly likely to screen. Censored probit predicts 72.0% of the unaware have a probability of screening greater than 0.5, while the sub-sample probit predicts this to be 83.5%.

The full-sample probit results lead to average probabilities that are much closer to the censored probit results. The kernel density plots for censored and full-sample probit illustrate the closeness for the unaware sub-group. Because there is only a modest amount of censoring the biases introduced by assuming all unaware women are non-screeners is less than that associated from ignoring the censoring problem. This is in accordance with our initial analysis of a stylised version of the problem that indicated full-sample probit is associated with small inconsistencies when censoring is minor and dependence is strongly positive. Both of these features are present in our data on screening behaviour in Australia.

Figure 1: Kernel density plots of estimated screening probabilities for the unaware sub-group
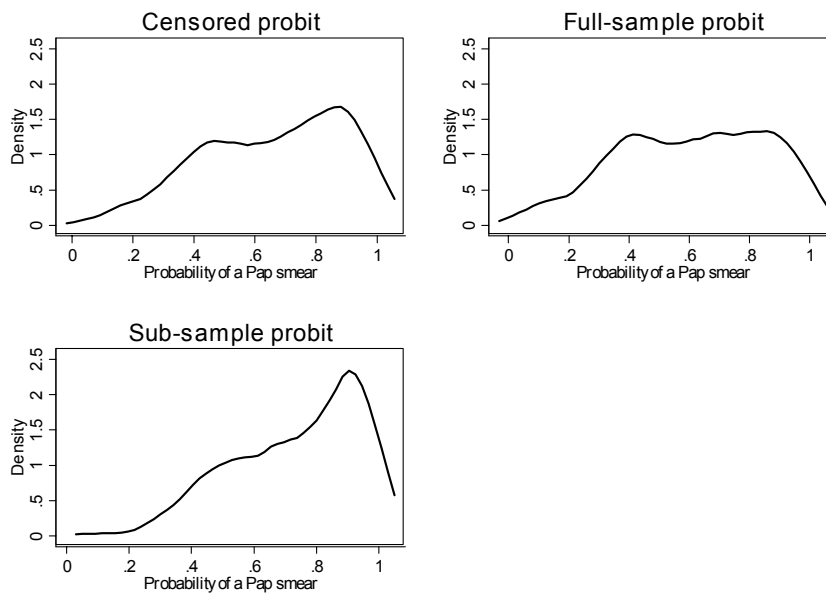
7. Conclusion

Our censored probit results provide a clear picture of Australian women's awareness of Pap tests and their propensity to ever screen for cervical cancer. While previous studies such as Taylor, Mamoon, Morrell and Wain (2001a,b) have recognised the role of age, socio-economic status and ethnicity in screening they have ignored issues associated with awareness and hence there is the potential for their results to be subject to the biases that we have discussed and illustrated.

The resultant modelling approach also allows us to gain new insights into how the key determinants impact on screening behaviour. For example, it is generally accepted that certain ethnic groups and women with poor language skills are over-represented amongst the under-screened. Our analysis has highlighted that a key part of the story with these women is that they are very likely to be unaware of Pap tests and hence their appearance as non-screeners is not necessarily due to any reticence to screen if they knew about the existence and availability of the tests.

From a methodological perspective we have emphasised the need to account for awareness in modelling choice. When the sample split into aware and unaware groups is unknown, one is forced to use full-sample probit. This may do a reasonable job of characterising choice probabilities, as it does in our particular case of screening, if

censoring is modest and positive dependence between awareness and choice is substantial. But there is no guarantee that such favourable conditions will be applicable in other cases. In fact for newer screening programs, such as colorectal cancer screening, it is likely that awareness will be much lower, and that there would be correlation between awareness and the choice to screen. If in fact censoring was modest and dependence was minor then full-sample probit would likely be inferior to sub-sample probit. But to implement the latter requires knowing who is aware and who is not.

Knowing the sample split allows one to make informed judgements about the appropriateness of alternative estimation procedures. Including questions in surveys that determine awareness is to be encouraged even when choices involve mature and well-known products or services as was the case here. Naturally, the incentives are even stronger when lack of awareness is likely to be widespread as in say a new test or procedure.

References

Ehrlich, I. and Becker, G. (1972), "Market insurance, self-insurance and self-protection", *Journal of Political Economy* 80, 623-649.

Greene, W.H. (2000), *Econometric Analysis*, (4th ed.), Macmillan Publishing Company, New Jersey.

Grossman, M. (1972), "On the concept of health capital and the demand for health", *Journal of Political Economy* 80, 223-255.

Hausman, J.A., Abrevaya, J. and Scott-Morton, F.M. (1998), "Misclassification of dependent variable in discrete-choice setting", *Journal of Econometrics* 87, 239-270.

Hsiao, C. and Sun, B.-H. (1999), "Modeling survey response bias – with an analysis of the demand for an advanced electronic device", *Journal of Econometrics* 89, 15-40.

Jones, A.M. (2000), "Health econometrics", Ch. 6 of A.J. Culyer and J.P. Newhouse eds., *Handbook of Health Economics*, North Holland.

Kenkel, D.S. (1990) "Consumer health information and the demand for medical care", *Review of Economics and Statistics* 72, 587-95.

Kenkel, D.S. (1991) "Health behaviour, health knowledge and schooling", *Journal of Political Economy* 99, 287-385.

Kenkel, D.S. (1994), "The demand for preventative medical care", *Applied Economics* 26, 313-25.

Kenkel, D.S. (2000), "Prevention", Ch. 31 of A.J. Culyer and J.P. Newhouse eds., *Handbook of Health Economics*, North Holland.

Leung, S.F. and Yu, S. (1996), "On the choice between sample selection and two-part models", *Journal of Econometrics* 72, 197-229.

Maddala, G.S. (1977), *Econometrics*, McGraw-Hill International Editions, Singapore.

Meng, C.L. and Schmidt, P. (1985), "On the cost of partial observability in the bivariate probit model", *International Economic Review* 26, 71-86.

Mullahy, J. (1999), "It'll only hurt a second? Microeconomic determinants of who gets flu shots", *Journal of Health Economics* 8, 9-24.

Nawata, K. and McAleer, M. (2001), "Size characteristics of tests for sample selection bias: A Monte Carlo comparison and empirical example", *Econometric Reviews* 20, 105-112.

Nelson, F. (1984), "Efficiency of the two-step estimator for models with endogenous sample selection", *Journal of Econometrics* 24, 181-196.

Taylor, Mamoon, Morrell and Wain (2001a), "Cervical screening in migrants to Australia ", *Australian and New Zealand Journal of Public Health* 25, 55-61.

Taylor, Mamoon, Morrell and Wain (2001b), "Cervical screening by socio-economic status in Australia", *Australian and New Zealand Journal of Public Health* 25, 256-260.

van de Ven, W.P.M.M. and van Praag, B.M.S. (1981), "The demand for deductibles in private health insurance", *Journal of Econometrics* 17, 229-252.

van Soest, A., Kooreman, P. and Kapteyn, A. (1993), "Coherency and regularity of demand systems with equality and inequality constraints", *Journal of Econometrics* 57, 161-188.

Zuehlke, T.W. and Zeman, A.R. (1991), "A comparison of two-stage estimators of censored regression models", *Review of Economics and Statistics* 73, 185-189.

Appendix 1: Inconsistency of alternative probit estimators

Our approach follows that used by van Soest, Kooreman and Kapteyn (1993) in a related context. In the simple two-equation system given in (2) there is a single explanatory variable x governing both awareness and screening and it is a dummy variable taking on a value of unity with probability 0.5. Three parameter values are set at particular values:

$$\beta_{11} = 0, \beta_{12} = 1, \beta_{22} = -1;$$

while two parameters are varied:

$$\beta_{21} = 1, 2, 3 \text{ and } \rho = -1.0, -0.8, \ldots, -0.2, 0, 0.2, \ldots, 0.8, 1.0.$$

Assuming β_{11} is known to be zero, the log-likelihood functions of sub-sample and full-sample probit will only be a function of β_{12}. For full-sample probit the (pseudo) log-likelihood is given by:

$$\log L(\beta_{12}) = \sum_i y_{1i} \log \Phi(x\beta_{12}) + \sum_i (1 - y_{1i}) \log[1 - \Phi(x\beta_{12})]$$

while the sub-sample probit counterpart will have the same structure but will involve likelihood contributions only from the subset of the observations that are not censored.

For any choice of (β_{21}, ρ), the population probabilities of possible outcomes can be determined and are defined as follows:

$\pi_{10} = \Pr[y_2 = 0 \mid x = 0];$ $\pi_{11} = \Pr[y_2 = 0 \mid x = 1];$
$\pi_{20} = \Pr[y_1 = 0, y_2 = 1 \mid x = 0];$ $\pi_{21} = \Pr[y_1 = 0, y_2 = 1 \mid x = 1];$
$\pi_{30} = \Pr[y_1 = 1, y_2 = 1 \mid x = 0];$ $\pi_{31} = \Pr[y_1 = 1, y_2 = 1 \mid x = 1].$

These are the probabilities that would enter the log-likelihood for the censored probit formulation that we have assumed is the true model and if estimated would produce consistent estimates of all unknown parameters. What is of interest is the behaviour of two single-equation probit models.

Let Π_{hk} be the number of observations associated with each of the $h = 1,2,3$ outcomes for each of the two values of $x = k$. If the total number of observations is $2N$, then:

$$p\lim_{N \to \infty}(\Pi_{hk} / N) = \pi_{hk}.$$

Using the true probabilities of each possible outcome, the probability limits of the (pseudo) log-likelihood functions of the mis-specified probit models can be determined. For full-sample probit this yields:

$$p \lim_{N \to \infty} \log L(\beta_{12}) = \frac{1}{2} \{\pi_{30} \log \Phi(0) + \pi_{31} \log \Phi(\beta_{12})$$

$$+ (\pi_{10} + \pi_{20}) \log[1 - \Phi(0)] + (\pi_{11} + \pi_{21}) \log[1 - \Phi(\beta_{12})]\}$$

$$= \frac{1}{2} \{\log(0.5) + \pi_{31} \log \Phi(\beta_{12}) + (\pi_{11} + \pi_{21}) \log[1 - \Phi(\beta_{12})]\}.$$

Because maximizing this function will give the same result as you would obtain from a simple binomial choice probit model, it is a simple matter to determine the probability limit of the resultant estimator of β_{12}. Suppose $\beta_{21} = 2$, and $\rho = 0.4$, then $\pi_{11} = 0.159$, $\pi_{21} = 0.105$, $\pi_{31} = 0.736$, and thus the required probability limit solves $\Phi(\beta_{12}) = 0.736$ which yields the full-sample probit entry of 0.63 given in Table 1 for this case.

Following the same approach, the (pseudo) log-likelihood function for sub-sample probit is given by:

$$p \lim_{N \to \infty} \log L(\beta_{12}) = \frac{1}{2} \{\pi_{30} \log \Phi(0) + \pi_{31} \log \Phi(\beta_{12}) + \pi_{20} \log[1 - \Phi(0)]$$

$$+ \pi_{21} \log[1 - \Phi(\beta_{12})]\}$$

$$= \frac{1}{2} \{(\pi_{20} + \pi_{30}) \log(0.5) + \pi_{31} \log \Phi(\beta_{12}) + \pi_{21} \log[1 - \Phi(\beta_{12})]\}.$$

Again this will produce the same estimates as a simple binomial choice probit model. Continuing the previous example, the required probability limit solves $\Phi(\beta_{12}) = 0.736/(0.736+0.105) = 0.875$ which yields the Table 1 entry of 1.15 for sub-sample probit.

Appendix 2: Extra estimation results

Table A1: Sub-sample and full-sample probit

Variables	Sub-sample Coefficient[a]	Standard Error	Full-sample Coefficient[a]	Standard Error
Constant	-0.4208*	0.2185	-1.1930***	0.1825
AGE2024	0.7615***	0.1052	0.7300***	0.0994
AGE2529	1.1749***	0.1166	1.1590***	0.1092
AGE3034	1.4057***	0.1325	1.3698***	0.1208
AGE3539	1.3630***	0.1350	1.3449***	0.1231
AGE4044	1.3827***	0.1345	1.3985***	0.1250
AGE4549	1.5599***	0.1332	1.5136***	0.1218
AGE5054	1.9117***	0.1525	1.8161***	0.1343
AGE5559	1.6042***	0.1385	1.4789***	0.1238
AGE6064	1.5159***	0.1358	1.4012***	0.1223
AGE6569	1.1313***	0.1241	1.0385***	0.1147
AGE7074	0.9221***	0.1244	0.8650***	0.1166
AGE7579	0.5972***	0.1325	0.5054***	0.1236
AGE80PL	0.3112**	0.1413	0.1240	0.1298
EXENG	-0.1555	0.1568	0.3598***	0.1155
ENGLISH	-0.0621	0.2093	0.4997***	0.1672
ENGEXENG	0.5195**	0.2265	0.0812	0.1837
NUBORN	-0.1353*	0.0732	-0.0792	0.0682
SEBORN	-0.1037	0.1280	-0.1242	0.1060
WEBORN	0.3776**	0.1864	0.3830**	0.1641
ASBORN	-0.4638***	0.1124	-0.6705***	0.0912
OTBORN	0.0134	0.1181	0.0446	0.1034
TERT	0.0923	0.0773	0.1386*	0.0711
DIPLOMA	0.0913	0.0860	0.0636	0.0760
TRADE	0.1088*	0.0637	0.1047*	0.0570
INC	0.0149***	0.0047	0.0132***	0.0043
INQSQ	-0.0002**	0.0001	-0.0001*	0.0001
INCMISS	0.3134***	0.0973	0.2704***	0.0865
SINC	-0.0004	0.0060	0.0039	0.0053
SINQSQ	0.0001	0.0001	0.0000	0.0001
SINCMISS	-0.0707	0.1332	0.0424	0.1157
VIC	-0.0763	0.0693	-0.0796	0.0612
SA	-0.1158	0.0751	-0.0596	0.0680
WA	-0.2143**	0.0857	-0.2013***	0.0768
TAS	0.1606	0.1202	0.1029	0.1060
NT	0.2297*	0.1194	0.0700	0.0987
ACT	-0.1090	0.0928	-0.0460	0.0844
BRISBANE	0.0320	0.1134	0.0600	0.1025
OTQLD	0.0056	0.1182	0.0556	0.1087
RURAL	-0.0732	0.0885	-0.0262	0.0814
REMRURAL	-0.0084	0.0723	-0.0243	0.0647
HPHI	-0.0031	0.0492	0.0486	0.0445
KIDS	0.6449***	0.0701	0.5910***	0.0613
SINGLE	-0.4041***	0.0940	-0.3280***	0.0811

Table A1: Continued

Variables	Sub-sample		Full-sample	
	Coefficient[a]	Standard Error	Coefficient[a]	Standard Error
SMOKE	0.2485***	0.0601	0.2794***	0.0547
QUIT	-0.0254	0.0719	-0.0572	0.0654
HLTHEX	0.0893	0.1162	0.0269	0.1048
HLTHVG	0.0877	0.1100	0.0881	0.0997
HLTHG	0.1908*	0.1098	0.0914	0.0987
HLTHF	0.1032	0.1164	0.0461	0.1042
Sample size	8580		8969	
Log-likelihood	-2029.26		-2588.26	
McFadden R^2	0.25		0.27	

Note: *** significant at 1% level, ** significant at 5% level, * significant at 10% level