Comparison of Active and Cancer Registry-based Follow-up for Breast Cancer in a Prospective Cohort Study

Ikuko Kato,1 Paolo Toniolo,1 Karen L. Koenig,1 Amy Kahn,2 Maria Schymura,2 and Anne Zeleniuch-Jacquotte1

The authors compared the relative effectiveness of two distinct follow-up designs in prospective cohort studies—the active approach, based on direct contact with study subjects, and the passive approach, based on record linkages with population-based cancer registries—utilizing available information from the New York University Women's Health Study (WHS) and the New York State Cancer Registry (NYSCR). The analyses were limited to breast cancer cases identified during the period 1985–1992, for which follow-up was considered reasonably complete by both the WHS and the NYSCR. Among 12,947 cohort members who reported a New York State address, 303 pathologically confirmed cases were identified through active follow-up and 284 through record linkage. Sixty-three percent of cancers were identified by both sources, 21% by the WHS only, and 16% by the NYSCR only. The agreement was appreciably better for invasive cancers. The percentage of cases identified only by the NYSCR was increased among subjects whose active follow-up was incomplete, as well as among nonwhites, obese patients, and parous patients. This suggests that relying on either type of follow-up alone may introduce certain biases in evaluating risk factors for breast cancer. Combining both approaches appears to be a better strategy in prospective cohort studies. Am J Epidemiol 1999;149:372–8.

In observational epidemiology, the prospective cohort study is a well established study design for testing of ecologic hypotheses (1). A disadvantage of these studies in comparison with case-control studies is the fact that, for diseases with low incidence, such as cancer, a large number of study subjects must usually be followed for an extended period of time in order to detect a sufficiently large number of cases of the disease. Because follow-up is an inherently labor-intensive and time-consuming activity, study costs are often an important consideration.

Choices of follow-up approaches will greatly affect the costs of prospective cohort studies. For example, in cancer epidemiology, follow-up costs could be reduced greatly by avoiding the direct follow-up of study subjects and relying entirely on record linkages with existing population-based registries. However, disadvantages of such an approach are that the clinical and pathologic information available from a registry is often limited, no personal information on study subjects is updated during the follow-up period, and it is not possible to take into account emigration of study subjects from the original study area.

The success of the studies based on passive follow-up depends in large part on the underlying assumption that all or most cases of disease in the study population will be detected by the registry. Conversely, studies based on active follow-up will be successful only with a low rate of loss to follow-up and high compliance of follow-up contacts, translating into identification of most outcome cases occurring during the study period. To our knowledge, no studies have compared the relative efficiency of passive and active methods of follow-up in prospective cohort studies. An excellent opportunity was offered by the New York University Women's Health Study (WHS), a prospective cohort study of mostly New York State residents.

MATERIALS AND METHODS

The New York University Women's Health Study

The WHS is a prospective cohort study designed to assess the role of endogenous hormones and diet in the etiology of breast cancer (2, 3). The cohort consisted
of 15,785 women aged 34–65 years who were undergoing mammographic screening at the Guttmann Breast Diagnostic Institute in New York City (N = 14,275) or the Strax Breast Cancer Institute in Florida (N = 1,510) between 1985 and 1991. Eligibility was limited to women who had not used steroid hormones, had not been pregnant, and had not breastfed a child in the 6 months preceding enrollment. At baseline, 82 percent of the study subjects and 90 percent of those recruited in New York reported a New York State address, and all participants provided essential information on risk factors for breast cancer in a self-administered questionnaire.

Active follow-up

After enrollment, all study members were actively followed to identify all incident cancers (except non-melanoma skin cancers) occurring in the cohort. Until 1991, active follow-up was based on self-administered questionnaires completed by the subjects returning to the clinic for annual screening. Questionnaires were also mailed to subjects who failed to return to the clinic for a period of 18 months or longer. After the end of recruitment, in 1991, questionnaires were mailed to all members of the cohort approximately every 18–24 months. In cases of no response, subjects were contacted up to three times. Subjects who had moved elsewhere were actively traced and contacted at the new address. The most recent follow-up effort for ascertainment of incident cases diagnosed before the end of 1994 began in December 1994 and has not yet been completed. Therefore, follow-up data for years beyond 1992 should be considered only partially complete.

With the consent of individual cohort members, cancer diagnoses reported on follow-up questionnaires were confirmed by obtaining a complete copy of pertinent medical records for review. Cancers were classified for tumor topography and morphology according to the clinical modification of the International Classification of Diseases, Ninth Revision (4).

The New York State Cancer Registry

The New York State Cancer Registry (NYSCR) was established in 1940, for the purpose of assessing the incidence of cancer within the state of New York, excluding New York City. In 1973, registration was extended to include New York City. The registry covers a population of more than 17 million. Cancer reporting is mandatory, according to state legislation. Reports are received from five sources: 1) physicians; 2) hospitals; 3) laboratories examining surgical specimens; 4) death certificates; and 5) reciprocal interstate reporting agreements. The completeness of reporting, based on the percentage of death-certificate-only registrations, has been estimated at 95 percent (5). The NYSCR followed International Agency for Research on Cancer rules for the registration of multiple primary cancers through 1995.

Record linkage

After appropriate institutional approval, the records of the 14,324 WHS cohort members who had a New York State address during the follow-up period or were recruited in New York or lost to follow-up were sent to the NYSCR. Personal identifiers included last and first names, date of birth, sex (all female), Social Security number, race, and last known address. Data on all identifiers except race (79 percent) and Social Security number (97 percent) were available for all subjects.

The cancer registry records used for this linkage contained cancers diagnosed between 1979 and 1995. At the time of our linkage, registration was considered complete as of the end of 1992 but incomplete afterwards. Linkage was performed using the six personal identifiers. Points were allocated to each exact match as follows: 10 points for Social Security number, seven points for complete date of birth, four points for the first three letters of the last name, four points for the first letter of the first name, one point for race, and zero points for sex (all study subjects were female). If dates of birth did not match exactly, two points were assigned for a matching year, month, or day, and one additional point was awarded if the birth years matched within 3 years. A complete match for sex and “Soundex” name (the first initial of the last name followed by a three-digit numeric code for the next three consonants, grouped by similarity of sound (6)) was mandatory (blocking variable). Definite matches were achieved with 18 or more points and possible matches with 14–17 points. These cutoff points were established through preliminary linkages so as to trade off sensitivity for specificity. After computer linkage, a meticulous review of all definite and possible matches was conducted with the use of additional information in the WHS databases, and 59 false matches were eliminated.

Statistical methods

The analyses were limited to incident cases of cancer occurring among 12,947 cohort members who had a New York State address. The percentages of all breast cancer cases identified by the NYSCR only, by the WHS only, and by both sources were calculated by year of diagnosis, type of cancer, and several epidemiologic variables. For the selected epidemiologic variables, the age-adjusted relative risks of breast cancer identified by the NYSCR or the WHS were estimated.
using Cox's proportional hazards model (7) among pre- and postmenopausal women separately. Menopausal status was determined at baseline. Ninety-five percent confidence intervals for the relative risks were obtained from the standard errors of the pertinent regression coefficients. Each subject's number of person-years at risk was computed as the amount of time from her baseline examination to the date of breast cancer diagnosis, death, or December 31, 1992, whichever occurred first. The capture-recapture method (8) was used to estimate the number of cases not identified by either procedure.

RESULTS

After exclusion of 42 records with unacceptable diagnoses and 368 cases diagnosed before the date of cohort enrollment, there were 762 NYSCR-linked records (1985-1995), while the WHS identified 792 incident cancers via active follow-up during the same period. Combining the two record series yielded a total of 1,014 incident cancers in this population. Of these, 540 were identified by both sources, 222 (21.9 percent) by the NYSCR alone, and 252 (24.9 percent) by the WHS alone.

Since more than 50 percent (n = 509) of the cases were cases of breast cancer and since our pathologic confirmation focused on this cancer, the analyses were limited to pathologically confirmed breast cancers (thereby excluding 29 cases without pathologic confirmation). In addition, since metachronous bilateral breast cancers were not registered as separate tumors by the NYSCR under the rules of the International Agency for Research on Cancer, the 23 bilateral cases were analyzed as 23 observations rather than 46.

Of the 457 breast cancer cases (table 1), 258 (56.5 percent) were identified by both sources, 88 (19.3 percent) by the NYSCR alone, and 111 (24.3 percent) by the WHS alone. When the data were examined by year of diagnosis, considerably more cases were identified by both approaches in earlier years than in more recent years, which probably reflects completeness of ascertainment for both procedures.

For breast cancer cases diagnosed between 1985 and the end of 1992, the period for which data from both the WHS and the NYSCR are assumed to be reasonably complete, concordance was 63 percent, with 21 percent of the cases being identified by the WHS only and 16 percent by the NYSCR only. After separating invasive and in-situ cancers (table 2), the percentage of cancers identified by the WHS only was higher for in-situ cancers.

When breast cancers diagnosed within 6 months of cohort enrollment were excluded from analysis (because the active follow-up took into account the results of baseline mammographic screening), the results were virtually unchanged: 61 percent of cases were identified by both sources, 16 percent by the NYSCR only, and 22 percent by the WHS only.

<p>| TABLE 1. Incident breast cancer cases identified by the New York State Cancer Registry (NYSCR), the New York University Women's Health Study (WHS), or both, by year of diagnosis: New York University Women's Health Study, 1985-1995 |</p>
<table>
<thead>
<tr>
<th>Year</th>
<th>NYSCR only</th>
<th>WHS only</th>
<th>Total no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1985-1987</td>
<td>10</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>1988-1990</td>
<td>24</td>
<td>30</td>
<td>54</td>
</tr>
<tr>
<td>1991-1992</td>
<td>24</td>
<td>27</td>
<td>51</td>
</tr>
<tr>
<td>1993-1995</td>
<td>30</td>
<td>34</td>
<td>64</td>
</tr>
<tr>
<td>Total</td>
<td>88</td>
<td>111</td>
<td>259</td>
</tr>
</tbody>
</table>

<p>| TABLE 2. Incident breast cancer cases identified by the New York State Cancer Registry (NYSCR), the New York University Women's Health Study (WHS), or both, by type of cancer: New York University Women's Health Study, 1985-1992 |</p>
<table>
<thead>
<tr>
<th>Cancer type</th>
<th>NYSCR only</th>
<th>WHS only</th>
<th>Total no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>All breast cancers</td>
<td>58</td>
<td>77</td>
<td>361</td>
</tr>
<tr>
<td>Invasive</td>
<td>50</td>
<td>61</td>
<td>303</td>
</tr>
<tr>
<td>In situ</td>
<td>8</td>
<td>16</td>
<td>58</td>
</tr>
</tbody>
</table>
Table 3 shows the numbers of cancer cases identified by the WHS, by the NYSCR, and by both sources, separately according to characteristics that may have influenced the success of the linkage and those that might be related to breast cancer risk. These analyses were also limited to cases diagnosed through the end of 1992. Among the subjects whose follow-up status was incomplete—i.e., those for whom the most recent follow-up information had been obtained before the end of 1992 and who were not known to be deceased—the percentage of cancers identified by the NYSCR only was strikingly higher (60.0 percent) than that among those whose follow-up status was complete (7.3 percent). Nonwhite cases were more likely to be identified by the NYSCR, while white cases were more likely to be identified by the WHS. Parous cases were more likely to be identified by the NYSCR than were nulliparous cases. In addition, the percentage of cases identified by the NYSCR only was higher among overweight cases (body mass index (weight (kg)/height (m)^2) ≥25) than among nonoverweight cases. Ascertainment did not vary appreciably according to age, residence outside of New York State, family history of breast cancer, or personal history of breast biopsy.

The relative risks of breast cancer were calculated for obesity (body mass index ≥25), parity (parous vs. nulliparous), and race (white vs. nonwhite), based on numbers of cases identified by the NYSCR (NYSCR only or both sources) or by the WHS (WHS only or both sources) (table 4). The relative risks based on either source were close to each other, and the differences ranged from 0.05 to 0.34. Results reflected those seen in table 3. A positive association with obesity in postmenopausal women tended to be more pronounced among NYSCR cases, and inverse associations with parity and obesity in premenopausal women tended to be so among WHS cases.
TABLE 4. Age-adjusted relative risk for breast cancer identified by either the New York State Cancer Registry (NYSCR) or the New York University Women's Health Study (WHS), according to obesity, parity, and race: New York University Women's Health Study, 1985–1992

<table>
<thead>
<tr>
<th>Menopausal status (at baseline)</th>
<th>Identification of cancer</th>
<th>NYSCR</th>
<th>95% CI</th>
<th>WHS</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parous</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body mass index £25.0</td>
<td>0.76</td>
<td>0.53–1.10</td>
<td>0.62</td>
<td>0.43–0.91</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>0.68</td>
<td>0.47–0.97</td>
<td>0.51</td>
<td>0.36–0.72</td>
<td></td>
</tr>
<tr>
<td>Parous</td>
<td>0.96</td>
<td>0.63–1.45</td>
<td>1.30</td>
<td>0.83–2.04</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body mass index ≥25.0</td>
<td>1.51</td>
<td>1.09–2.10</td>
<td>1.32</td>
<td>0.97–1.80</td>
<td></td>
</tr>
<tr>
<td>Parous</td>
<td>0.95</td>
<td>0.66–1.38</td>
<td>0.82</td>
<td>0.58–1.15</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>0.54</td>
<td>0.34–0.83</td>
<td>0.59</td>
<td>0.38–0.91</td>
<td></td>
</tr>
</tbody>
</table>

* RR, relative risk; CI, confidence interval.
† Weight (kg)/height (m²).

DISCUSSION

Our study suggests that, in regions that are served by a population-based cancer registry, a satisfactory level of case ascertainment can be achieved in prospective cohort studies by combining active follow-up and record linkage with the cancer registry. Utilizing a capture-recapture approach (8), we estimated that 95 percent of breast cancer cases in our population were ascertained by combining both follow-up approaches. We also estimated that approximately 20 percent of breast cancer cases would have been missed if we had relied on only one of these two types of follow-up.

The better ascertainment of breast cancer by active follow-up in our population may not be surprising, since all subjects were enrolled in a breast cancer screening program and had volunteered to participate in breast cancer research. Breast cancer was their concern, and they might have been especially motivated to report their breast cancer experience promptly and correctly. Furthermore, the validity of self-reported breast cancer has been reported to be excellent (9), although this may apply only to highly selected populations.

The percentage of cancers identified only by the NYSCR increased in recent follow-up periods, which suggests that the completeness of our active follow-up may have dropped over time. This may reflect changes in the WHS follow-up procedures. After 1987, we began reducing the size of WHS staff in charge of recruiting new participants and contacting old participants for follow-up purposes at the screening clinic. At that time, direct contacts made at the clinic were replaced by indirect contacts, made primarily by mail. Thus, a reduction in the intensity of active follow-up may have been reflected in the gradual decrease in case ascertainment by our follow-up. However, our data suggest that most of the subjects with incomplete follow-up resulting from changes in follow-up complexity were later identified by linkage with the cancer registry.

The availability and accuracy of personal identifiers in both the registry and the cohort and the computerized matching algorithm in use are other important determinants of successful linkage. The NYSCR matching algorithm utilized weighting for each matching criterion (Social Security number, date of birth, last and first names, and race) in determining definite and possible matches. A similar algorithm has been used in record linkage with the National Death Index Search, as well as in other population-based cancer registries and surveys, and has proven to be very efficient (10–14). However, a limitation is that the NYSCR weighting system is rather simplified, neither value-specific nor exactly probabilistic. The authors recognize that a standardized computerized linkage method, Automatch (Matchware Technologies, Inc., Burtonsville, Maryland) (15), which is based on probabilistic methods that weight information fields, following work published by Fellegi and Sunter (16), has recently become available, and some cancer registries have begun to introduce it. The introduction of such a program in the future may be useful in validating the results of the present study.

The sensitivity of our linkage can be calculated as 75 percent among 303 breast cancer cases with pathologic confirmation identified by WHS through the end of 1992. If nine cases without pathologic confirmation identified by the NYSCR are considered matches, the sensitivity increases to 78 percent. Subsequent to the linkage, an intensive search of the NYSCR database was undertaken for cases that had not been identified by the matching algorithm. This search included 57 breast cancer cases, excluding the 11 subjects with a
non-New York State address and the nine cases registered without histologic confirmation from the total of 77 subjects identified by the WHS only in table 2. Among these 57 cases, 11 cases were located. The major reasons for the unsuccessful linkage were distinct differences or misspellings in surnames, sufficient to result in different Soundex coding, and apparent differences in dates of birth. For the remaining 46 cases, medical record information in the WHS proved that two were diagnosed at hospitals outside of New York State or the United States but 44 were diagnosed within New York State. In summary, among the 77 cases, 14 percent were missed because of the matching algorithm, 17 percent because of migration, and 69 percent because of incomplete registration. Consequently, the sensitivity of the linkage algorithm would be calculated to be 95 percent, excluding the 66 cases from the denominator.

The specificity of the linkage algorithm was estimated to be 99.5 percent, based on the fact that there were 59 false-positive matches among 12,577 persons for whom no cancer had ever been reported by either method of follow-up. We realize that approximately 7 percent of the incident cancer cases were not captured by either procedure, according to the capture-recapture analysis (8). However, if one assumes that these cases were similarly distributed among the subjects with and without a false match, our estimate of the specificity can be justified. It should also be noted that it was not feasible to assess specificity on the basis of false-positive diagnoses, because this investigation was limited to cases with histologic confirmation.

It may not be appropriate to generalize our experience with the WHS to other prospective cohort studies. The success of record linkage between a prospective cohort study and a cancer registry is likely to be a function of the method used and the completeness of active follow-up, as well as the quality, completeness, and size of the cancer registry and the mobility of the population. It is also likely to be a function of the specific characteristics of the study population. For example, our study included mostly middle-aged, middle-class women, with a relatively small representation of ethnic minorities (16 percent). The potential for success of the linkage in a study including only men or only minority subjects could hardly be inferred on the basis of our data. Record linkage may be more efficient in male populations, since women are more likely to develop cancer, and size of the cancer registry increases. Therefore, the efficiency of long-term follow-up seems to depend on the characteristics of the study population.

Incomplete identification of cases may be sufficient to bias the results of epidemiologic studies. When such a possibility was evaluated by calculating relative risks according to selected risk factors for breast cancer, the difference in the risk estimates based on NYSCR and WHS cases were generally small, although it could be greater for some strata. The bias should be minimal provided that either a fairly high proportion of cases is identified by both sources or the risk factor is similarly distributed in the groups identified by only one source.

In summary, our study indicates that in areas served by a good-quality population-based tumor registry, such as New York State, prospective cohort studies relying exclusively on tumor registry data for subjects’ follow-up may fail to identify a substantial number of incident cancer cases arising in the study population. Conversely, prospective cohort studies relying exclusively on their own, active follow-up may also fail to identify a substantial percentage of their cancer cases. When it is economically unfeasible to combine passive and active follow-up, it seems important that investigators be aware of the problem and at least attempt to assess its dimensions in their own study populations.

ACKNOWLEDGMENTS

This study was supported by grants CA34588 and CA16087 from the National Cancer Institute.

REFERENCES