Published in final edited form as:
Oncol Rev. 2011 September 1; 5(3): 143–147. doi:10.1007/s12156-011-0077-0.

A limited Review of Over Diagnosis Methods and Long Term Effects in Breast Cancer Screening

Dongfeng Wu[1] and **Adriana Pérez**[2,*]

[1] Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville, Louisville, KY 40202, USA

[2] Division of Biostatistics, School of Public Health, Austin Regional Campus, The University of Texas Health Science Center at Houston and Michael & Susan Dell Center for Healthy Living, Austin, TX, USA

Abstract

Breast cancer screening programs have been effective in detecting tumors prior to symptoms. Recently, there has been concern over the issue of over-diagnosis, that is, diagnosis of a breast cancer that does not manifest prior to death. Estimates for over-diagnosis vary, ranging from 7% to 52%. This variability may be due partially to issues associated with bias and/or incorrect inferences associated with the lack of probability modeling. A critical issue is how to evaluate the long-term effects due to continued screening. Participants in a periodic screening program can be classified into four mutually exclusive groups depending on whether individuals are diagnosed and whether their symptoms appear prior to death: True-early-detection; No-early-detection; Over-diagnosis; and Not-so-necessary. All initially superficially healthy people will eventually fall into one of these four categories. This manuscript reviews the major methodologies associated with the over-diagnosis and long-term effects of breast cancer screening.

Keywords

true early detection; no early detection; long term effect; probability modeling

1. Introduction

Breast cancer screening programs have been effective in detecting early stage tumors prior to symptoms. One unresolved problem that is associated with regular screening is the evaluation of over-diagnosis, the diagnosis of “disease” that will never cause symptoms or death during a patient’s lifetime [1]. Another problem is how to evaluate the long-term effects due to continued regular screening. For example, how should the probability of over-diagnosis and true-early-detection be quantified in regular screening? Will regular screening exams result in a greater chance of over-diagnosis?

To date, the majority of research that has been done in the area of over-diagnosis has been based on observational studies [2–8], rather than on probability modeling [9–10]. The estimated percentage of over-diagnosis in breast cancer screening varies widely from 7% [2] to 52% [3]. Hence, there is controversy concerning the benefit of continued regular

*Author to whom correspondence should be addressed; Tel.: +1-512-391-2524; Fax: +1-512-482-3185. adriana.perez@uth.tmc.edu.

Conflict of interest statement
The authors declare that they have no conflict of interest.

screening. The following review considers the four primary methods that have been used to assess over-diagnosis associated with cancer screening. The first two methods were based on observational studies meanwhile the last two methods were based on probability modeling.

2. Comparing incidence rates in the screening group and the control group

Zackrisson et al. [2] provided a simple way to estimate the rate of over-diagnosis. They compared the incidence rates in the screening group and the control group in the Malmö, Sweden mammographic screening trial. The population they studied was 42,283 women aged 45–69; they were randomized to two groups: screening group and control group (no screening), with about 21,000 in each group. The trial was carried out during 1976 and 1986, with a long follow up until 2001. They used Cox's proportional hazards analysis to estimate relative incidence rates in both groups based on the 15-years follow up data. Fifteen years passed after the end of the trial, the screened group (46.6%) has a 10% higher incidence of breast cancer than their control group (36.3%), including both invasive and in situ cancer. That rate is 7% higher in the screened group if cancer in situ was excluded. The statistical analysis they used was simple, but makes sense. They argued that the over-diagnosis should be assessed by studying the cumulative incidence of breast cancer over time in the two groups at the same time period.

3. Estimating relative incidence of over-diagnosis after adjustment

Duffy et al. [5] provided a proposal to estimate over-diagnosis in breast cancer after adjusting for three complex issues. The first issue is the need for removing the prevalence of screening of those women reaching the lower age limit that coincides with the starting of any introduction of a screening program in a particular population. These authors encourage researchers to identify and separate incidence screening from prevalence screening. Therefore, an estimate of incidence can be made excluding prevalence screening. The second issue is the need to adjust the age specific incidence due to lead time. For this adjustment researches need to provide an estimate of averaged lead time on screen detected cancer. The third issue is the need to adjust the incidence due to any changes in incidence (i.e. anticipation of the temporal trend or treatment intervention effects).

These authors used the breast cancer incidence in Sweden ages 50–69 to apply their proposal. The target population for screening in 11 counties in Sweden was 463,405 women and the breast cancer relative to that expected from prescreening trends was 1.84. After adjustment of the prevalence screening, the incidence was estimated to be 1.69. After estimating a 2.4 year lead time, then the estimate of incidence was reduced to 1.49. After adjusting for changes due to the use of hormone replacement therapy (HRT), then the adjusted relative incidence reports by these authors as a rough approximation was 1.39, "a 39% excess." However, the authors claimed that this is not the correct estimate of over-diagnosis; but it helps researchers in understanding the complexity of estimation of over-diagnosis [5].

4. Evaluating schedule over-diagnosis using probability modeling

Davidov and Zelen [9] introduced two measurements of over-diagnosis in a fixed screening program: individual over-diagnosis and schedule over-diagnosis. Assume that there will be m scheduled exams $\tau = \tau, \cdots, \tau_m$) and assume that cancer symptoms develops through three states: disease-free, pre-clinical and clinical, they used the forward recurrence time to derive the conditional probability of over-diagnosis for an i-th generation person who was diagnosed at the j-th exam (j ≥ i) at age τ_j, which they called the individual over-diagnosis. Then, they summarized these probabilities to get the conditional probability of over-diagnosis at all exams for this fixed screening schedules, which they called the schedule

over-diagnosis. This is the over-diagnosis that most of researchers are interested in. They applied their method to male prostate cancer screening and estimated the schedule over-diagnosis. Their estimate depends on the screening sensitivity, the sojourn time in the preclinical state, and the transition probability density from the disease-free to the preclinical state. In their simulation, they assumed that sensitivity is 0.3, 0.7, and 0.9; the sojourn time follows an exponential distribution, with a mean sojourn time of 5, 7.5, 10, 12.5, and 15 years. For the transition probability density, they estimated it from the Surveillance Epidemiology and End Results (SEER) database first, and then substituted the result into their model. For the probability of death within one year of a person at age k, they used the published Period Life Table from the Social Security Administration for the US males in 1997. They simulated 3 different exam schedules, τ = (50, 55, 60), (50, 55, 60, 65, 70), and (50, 55, 60, 65, 70, 75, 80), that is, screening was carried out every 5 years, but with different numbers. See Table 1 in Davidov and Zelen [9].

They found that the risk of over-diagnosis for prostate cancer was changing between 8% and 52%, depending on different assumptions. The probability of over-diagnosis is barely changing when the screening sensitivity changes from 0.3 to 0.9; however, it increases dramatically when the sojourn time increases from 5 years to 15 years. If the sojourn time is 5 years, the probability of over-diagnosis will be around 23% for the longest exam schedule; and it will jump to 52% if the sojourn time is 15 years long. Their simulation results suggested that the sojourn time plays the most important role in the possibility of over-diagnosis.

Although the simulation for this paper was for prostate cancer, the method can be applied to breast cancer as well. The key finding is that the probability of over-diagnosis depends on sojourn time mainly. Since breast cancer has a much shorter sojourn time (from one to four years) than that of prostate cancer, the risk of over-diagnosis should be much smaller. Their method has a few limitations. For example, the model can only handle the situation when the total number of screening exams is fixed beforehand. In another words, this model is retrospective: when the screening was done, and we looked back, we can figure out the probability of over-diagnosis. So the lifetime is not really a random variable, but a fixed upper bound, such as 60, 70, or 80 in their simulation. Their derivation is quite complicated and could be greatly simplified.

5. Probability modeling to long term effects, including over-diagnosis

Instead of dealing with over-diagnosis alone, Wu and Rosner [10] addressed the long-term effect for the whole cohort, using probability modeling. Initially superficially healthy people in a periodic screening program can be classified into four mutually exclusive groups: True-early-detection, No-early-detection, Over-diagnosis, and Not-so-necessary. Eventually every initially superficially healthy participant will fall into one of the four groups. The assumption of this disease progressive model is that: everyone who has cancer will develop through three states: the disease-free, the pre-clinical, and the clinical state. Another assumption is that an individual is asymptomatic and without a history of breast cancer before she takes her first breast cancer screening exam. These are common assumptions as in Davidov and Zelen [9]. However, the major difference is that the life time in Wu and Rosner [10] is really a random variable, hence the number of screening exam is not a fixed number, but a random variable as well. So it is a projection to the future when someone decides to pick a schedule. Based on a woman's diagnosis status and her ultimate lifetime disease status, Wu and Rosner [10] categorized people who take part in periodic screening into four mutually exclusive groups as following.

- Case 1: Not-so-necessary (NsN). A woman who took part in screening exams that never found breast cancer and ultimately died of other causes.

- Case 2: No-early-detection (NoED). A woman who took part in screening exams, but whose disease manifest itself clinically and was not detected by screening.
- Case 3: True-early-detection (TrueED). A woman whose breast cancer was diagnosed at a scheduled screening exam and whose clinical symptoms would have appeared before her death.
- Case 4: Over-diagnosis (OverD). A woman who was diagnosed with breast cancer at a scheduled screening exam but whose clinical symptoms would NOT have appeared before her death.

One of the presented authors first derived the probability for each case when the number of screening is fixed, then we allow the number of screening to be a random variable, because we treated the human lifetime as a random variable and people can die of other causes besides breast cancer. The probability of these four cases was in fact a function of the screening sensitivity, the sojourn time distribution, the transition probability density from disease free to the preclinical state, the screening frequency and age at first screening.

Then we applied their method to the Health Insurance Plan of great New Yorker (HIP) breast cancer screening data [11]. We first get the information of the sensitivity, the transition probability density and the sojourn time distribution from the HIP data by using the likelihood model we developed with a Bayesian inference. These three key parameters were reported through the 2000 Bayesian posterior samples [12]. Then we substitute in these 2000 posterior samples into the four probability formulae to estimate the probabilities of each category under different screening frequencies and different initial ages. The lifetime in the simulation is a random variable, and they derived the lifetime density function from the actuarial life table of the Social Security Administration that first published in 2006 [13]. In these simulations, the initial screening ages were 40, 50, and 60; the screening intervals were 6, 12, 18, 24 and 30 months. We summarized projected probabilities for breast cancer screening using HIP study data in Table 1.

In Table 1, for all three age groups, the probability of not-so-necessary is very high; it increases from about 89% to 93% when the initial screening age increases from age 40 to 60; it has little changes when the screening frequency changes within each age group. The commonly perceived lifetime risk for breast cancer is 1 in 9 or 11%, and most of the breast cancer cases would happen after age 40. Their estimated probability of not-so-necessary is about 89% for the 40-year-old group, or it equals 1 minus the perceived lifetime risk, which is compatible to our common sense. Researchers interested in the estimation of the proportion of all cancers detected can estimate this by adding columns 3, 4 and 5 in table 1. Table 1 also shows the overall percentage of over-diagnosis is very low, less than 1% among all participants. Please notice that the results presented in table 1 are not the conditional probabilities, but the probabilities for the whole HIP cohort. These results are different from the probability discussed by other researchers who report the conditional probability among screen detected cases. For this reason, we summarized the projected probabilities of true-early detection and over-diagnosis for screen-detected cases in Table 2.

Table 2 showed the probability of true-early-detection and the probability of over-diagnosis among those whose cancer would be diagnosed by regular screening exam. This definition of probability is the commonly discussed probability of over-diagnosis. It is about 5~7%; and the probability of true-early-detection is very high, above 93~95% among those detected by screening. This result is closer to that of Zackrisson et al [2] found in their observational study. Since the HIP study was the first mass screening study and it was carried out in the 1960s, the screening sensitivity was lower at that time; the cancer incidence was also slightly lower at that time. This method, though not perfect, can serve as a framework to evaluate the long-term effects of cancer screening for the whole cohort.

6. Conclusion and Discussion

Duffy et al [5] showed how complicated it is to estimate the rate of over-diagnosis in breast cancer screening, because of the lead time bias and other factors involved. They pointed out that using the breast cancer incidence to make inference for over-diagnosis has many drawbacks, for example, the increased HRT use probably caused much higher breast cancer incidence in the 1990s, and their informal adjustments may still over-estimate the percentage of over-diagnosis. Jorgensen and Gotzsche [3] compared the trends of over-diagnosis before and after screening. They estimated over-diagnosis using linear regression as 52%, with a 95% confidence interval of (46%, 58%). However, their estimates do not take into account other factors, e.g. lead time bias, so such estimates of over diagnosis rate may be severely over-estimated.

We reviewed two typical methods using observational studies in section 2 and 3. Though both methods did not used probability models, the first one seems more reliable and it makes sense in their estimation of over-diagnosis because they were comparing the incidences during the same time period. However, researchers planning to use the method described in section 2 need the cumulative incidence data on a screening group and a corresponding control group with a long follow-up period. This may not be available. Also, their estimation is only applicable to that specific study, and cannot be extended to other screening cohort with a different screening frequency. This is because we know, the proportion of over-diagnosis is changing with screening frequency.

We also reviewed two different probability models in section 4 and 5. Davidov and Zelen [9] model was the first probability model, though it has many limitations, such as the lifetime really has a fixed upper bound, and their estimated probability of over-diagnosis is changing with this upper bound. Their simulation is for prostate cancer; however, their method is a contribution to this area, which is why it is included in this review. Wu and Rosner [10] developed a new model to evaluate the whole cohort of the screening program by categorizing all initially superficially healthy people into four mutually exclusive categories: true-early-detection, over-diagnosis, no-early-detection and not-so-necessary. The advantage is that it provides a systematic approach to evaluate the long term effects of the screening program. The method is also a predictive study: using estimations from the sensitivity, the sojourn time distribution, and the transition probability, the method can estimate the probability of each of the four categories under different screening frequency, and different initial age at screening.

We have selected some methodologies for this review with the hope that the readers will realize the challenges of the problem of over-diagnosis in breast cancer screening studies.

Acknowledgments

The authors thank Dr. Kathy Baumgartner for her valuable comments and suggestions. We thank the anonymous reviewers for their valuable suggestions and comments for improving this paper. The second author was supported by a research supplement (3R37CA057030-20S1) from the National Cancer Institute during the writing of this manuscript.

References

1. Day NE. Overdiagnosis and breast cancer screening. Breast Cancer Research. 2005; 7:228–229. [PubMed: 16168144]
2. Zackrisson S, Andersson I, Janzon L, Manjer J, Garne JP. Rate of over-diagnosis of breast cancer 15 years after end of Malmö mammographic screening trial: follow-up study. BMJ. 2006; 332:689–692. [PubMed: 16517548]

3. Jørgensen KJ, Gøtzsche PC. Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends. British Medical Journal. 200910.1136/bmj.b2587

4. Badgwell BD, Giordano SH, Duan ZZ, et al. Mammography before diagnosis among women age 80 years and older with breast cancer. Journal of Clinical Oncology. 2008; 26:2482–2488. [PubMed: 18427152]

5. Duffy SW, Lynge E, Jonsson H, et al. Complexities in the estimation of overdiagnosis in breast cancer screening. British Journal of Cancer. 2008; 99:1176–1178. [PubMed: 18766185]

6. Gøtzsche PC, Jørgensen KJ, Mahlen J, Zahl PH. Estimation of lead time and overdiagnosis in breast cancer screening. British Journal of Cancer. 2009; 100:219. [PubMed: 19127274]

7. Zahl PH, Mahlen J, Welch HG. The natural history of invasive breast cancers detected by screening mammography. Arch internal Medicine. 2008; 168:2311–2316.

8. Paci E, Duffy SW. Overdiagnosis and overtreatment of breast cancer: overdiagnosis and overtreatment in service screening. Breast Cancer Research. 2005; 7:266–270. [PubMed: 16457702]

9. Davidov O, Zelen M. Overdiagnosis in early detection programs. Biostatistics. 2004; 5:603–613. [PubMed: 15475422]

10. Wu, D.; Rosner, GL. Proceedings of the American Statistical Association,2010. Biopharmaceutical Section. Alexandria, VA: American Statistical Association; A projection of true-early-detection, no-early-detection, over-diagnosis and not-so-necessary probabilities in tumor screening; p. 1144-1157.

11. Shapiro, S.; Venet, W.; Strax, P.; Venet, L. Periodic Screening for Breast Cancer: The Health Insurance Plan Project and its Sequelae, 1963–1986. The Johns Hopkins University Press; Baltimore, USA: 1988.

12. Wu D, Rosner GL, Broemeling LD. MLE and Bayesian inference of age-dependent sensitivity and transition probability in periodic screening. Biometrics. 2005; 61:1056–1063. [PubMed: 16401279]

13. Social Security Administration. [Accessed 01 August, 2010.] Period Life Table. Actuarial Publications, 2006. http://www.ssa.gov/OACT/STATS/table4c6.html

Table 1

The projected probability (%) for the whole cohort using HIP study data

Screening interval	P(NsN)	P(No-ED)	P(True-ED)	P(Over-D)
Age at initial screen = 40				
12 mo.	89.71	2.48	7.34	0.32
24 mo.	89.78	4.77	5.04	0.25
Age at initial screen = 50				
12 mo.	91.21	1.95	6.49	0.33
24 mo.	91.28	3.91	4.52	0.26
Age at initial screen = 60				
12 mo.	93.21	1.39	5.03	0.33
24 mo.	93.28	2.86	3.57	0.26

Source: Wu, D and Rosner, G.L. A projection of true-early-detection, no-early-detection, over-diagnosis and not-so-necessary probabilities in tumor screening. Proceedings of the American Statistical Association,2010. Biopharmaceutical Section, Alexandria, VA: American Statistical Association. 1144–1157.

Table 2

The projected probability of true-early-detection and over-diagnosis for screen-detected cases

	Initial Age 40		Initial age 50		Initial age 60	
Screening interval	**P(True-ED)**	**P(Over-D)**	**P(True-ED)**	**P(Over-D)**	**P(True-ED)**	**P(Over-D)**
12 mo.	95.47%	4.53%	94.84%	5.16%	93.43%	6.57%
24 mo.	95.12%	4.88%	94.50%	5.50%	93.10%	6.90%

Source: Wu, D and Rosner, G.L. A projection of true-early-detection, no-early-detection, over-diagnosis and not-so-necessary probabilities in tumor screening. Proceedings of the American Statistical Association,2010. Biopharmaceutical Section, Alexandria, VA: American Statistical Association. 1144–1157.