Arch Public Health
2000, 58, 131-144

Evaluation of a pilot quality assurance programme for breast cancer screening in the Brussels area

by

Renard F. [1], Bourdon C.D. [2], Andry M. [1], Mendez V. [2], Grivegnée A-R. [1], Vandenbroucke A. [2]

Abstract

Background: *a pilot quality assurance project for breast cancer screening was set up in the Brussels area in 1994. This paper aims to assess the performance of this programme after 4 years of activity, and the specific impact of consensual double reading of mammograms.*

Methods: *each screening mammogram of women aged 50-69 year was submitted to a consensual double reading. Results of readings were registered with standardised forms. Follow-up data were traced for every positive mammogram.*

[1] Jules Bordet Institute, Epidemiology Unit, Brussels.
[2] Catholic University of Louvain, Cancer Prevention Unit.
Address correspondence to :
Dr Françoise Renard, Centre de Référence pour le Dépistage du Cancer du sein, 479 chaussée de Louvain, 1030 Brussels. Tel: 32 2 742 21 34 or 32 2 541 30 59, Fax: 32 2 742 21 33, E-mail: crdcs@beon.be

This work have been supported by the "Europe against Cancer" project, the French Community of Belgium, the French Commission of Brussels-Capital Region and the Federation against Cancer.

Results: *15.624 mammograms were performed in 12.239 women; recall rate at first round was 7,8%, open biopsy rate was 1%, cancer detection rate was 5,8‰, positive predictive value of biopsy recommendation was 53,4%, benign to malignant biopsy ratio was 0,87:1, small size (less or equal to 10 mm) cancer proportion was 40%, proportion of cancers free of nodal involvement was 65%. Double reading yielded a 6% gain in sensitivity, while recall rate dropped from 8,1% to 7,8%.*

Conclusion*: apart from a too high recall rate, screening performance was comparable with other published results in the same context; performance indicators ranged within norms recommended by "Europe against cancer". However, impact of double reading was weak and should be re-evaluated in the future perspective of a larger scale organised programme.*

Key-words

Quality assurance, mammography screening, double reading, early indicators, evaluation.

Introduction

Randomised studies have shown that breast cancer screening has the potential to reduce mortality from breast cancer at least in women aged 50-69 years (1-5). To achieve some reduction in mortality, two conditions are required: sufficient attendance rate in a breast cancer screening programme (at least 70%), and a high quality screening. Organised screening programmes, including a population-based invitation of women and quality assurance procedures are the best strategy to reach conditions of efficacy. The set up of such organised breast cancer screening programmes in all European countries has been recommended by the programme "Europe against cancer" since the early nineties (6). Those programmes should usefully be prepared by local pilot projects, which should then be extended to national or regional programmes.

Belgium was late to follow those recommendations: until 1999, only small pilot projects have been set up. However, since Belgium is a federal country where preventive public health policies are decided at a

regional level, the situation is quite different in the North and the South of the country. The Government of the Flemish Community supported pilot projects since the middle of the nineties. In 1999, after the results of the evaluation of those projects (7), the Flemish Government decided to set-up a regional programme in the whole Flanders. In opposition, in Brussels and Wallonia, there is not yet a governmental decision to set-up such a programme, and the pilot projects are sparsely supported by the regional government. Therefore, pilot projects in Brussels and Wallonia are until now mostly restricted to the quality assurance aspects, at the exception of one project developed in the suburbs of Liege (8). Discussion between the team leaders of the pilot projects and the regional government have now begun, in hope to obtain a decision to set-up a fully organised programme in Brussels and Wallonia.

The "Brussels Project for Breast Cancer Screening" is actually part of the "Reference Centre for Breast Cancer Screening", which is a quality assurance project running in Brussels and some part of Wallonia (actually, Walloon Brabant and locality of Dinant). The project focuses on setting up quality assurance procedures in existing mammographic facilities. Women are not invited and screening is then a self-referred or physician-referred process. This quality-assurance-focused phase should be the preliminary phase of a full organised programme with a systematic invitation of women.

Quality assurance procedures involve technical assessment of equipment and imaging, training of radiologists, double reading of mammograms and audit of performance results. Sets of standardised performance indicators have been developed by several authors (9-11) for this assessment, addressing the *quality* of the screening process and its *efficacy*, assessed by prognostic indicators of the screen-detected cancers. Some of those indicators are recommended by the programme "Europe against Cancer" (12), which has also advised target values to obtain.

Pilot projects should be considered as a necessary preparatory phase for a regional project. They should be carefully evaluated before extension on a larger scale.

This study aims to assess performance results of the Brussels quality assurance programme for breast cancer screening during it's first 4 years of running.

Materials and methods

Description of the project

At the beginning of the Project (June 1994), 7 radiologic units were involved. This number rose to 9 in 1995, all situated at Brussels. The mammographic equipment have to satisfy to European norms; it is initially assessed ("acceptance test") and is submitted to six-months controls, and daily calibration (13). Participating radiologists sign an agreement with the project whereby they agree to perform the quality controls of the equipment and to partially fund it, to submit all screening mammograms of women aged 50-69 years to the project for double reading, to use a standardised registration form for those screening mammograms, and to perform themselves a double reading session once a week.

The first reading of mammograms is performed at the peripheral radiologic units. The Project centralise all screening mammograms of women aged 50-69 years for double reading and the corresponding forms for registration. Double reading is performed at the Project by one radiologist of the pool of radiologists having performed first reading; of course, mammograms are never read twice by the same radiologist. In case of discrepancy, a consensual third reading is performed, in the presence of both radiologists which performed the first and the second reading, and of a more experienced radiologist.

Type and source of data

Mammographic results on standardised forms have been obtained from all participating centres. Registration was estimated to be carried-out in 85% of mammograms performed; loss was due to organisational incidents, unrelated to the result of mammograms. Follow-up information for screen-detected abnormalities was actively searched by sending to radiologists and physicians a follow-up form.

Results of mammograms were classified as "normal" or "abnormal", meaning that the women had to undergo further workup (additional imaging, fine needle aspiration or biopsy) or to repeat mammogram in 6 months. Following indicators were calculated from mammographic forms: recall rate at initial reading (defined as the proportion of abnormal mammograms), recall rate according to final conclusion, concordance between readers. Concordance was measured by the kappa coefficient,

which corrects observed concordance for random effects. As kappa coefficient alone is unable to resume all aspects of concordance and is affected by prevalence, sensitivity and specificity of the test, two additionnal concordance indices are recommended (14). Positive agreement is defined as the proportion of agreement from the average number of positive conclusion at any reading. Negative agreement was the proportion of agreement from the average number of negative conclusion at any reading

Follow-up forms sent to radiologists and/or to practitioners collected data on biopsies and their results: malign/benign character of the lesion, behaviour of the tumour, tumour size, nodal involvement. Usual assessment of the programme performance involved measure of recall rate, detection rate, positive predictive value for biopsy indication (number of cancers within biopsies), biopsy rate, benign to malignant biopsy ratio, percentage of small size invasive tumours (less or equal to 10 mm), percentage of nodal free malignant tumours, approximated specificity (defined (10) as: number of negative screening tests divided by total number of tests minus true positive number). Prevalence/expected incidence ratio was calculated as the ratio between detection rate in our program and yearly age-specific incidence recorded at the National Cancer Register, for the whole country in 1995 (latest available figures). Improvement in detection rate yielded by the whole double reading process was calculated as the mean number of cancers detected by only one reader divided by number of cancers detected by both readers plus mean number of cancers detected by only one reader (15, 16). As the link with the National Cancer registry is not operational yet in Belgium, sensitivity of the test could not be assessed.

This whole set of performance indicators was calculated for all the 9 mammographic facilities as a whole, at first round and at subsequent screening (next rounds screening).

Results

During the study period (from june 1994 to april 1998), 15.624 mammograms were sent by the 9 participating mammographic facilities to the double reading centre. 12.239 mammograms were first round screening mammograms in the programme, and 3.385 were subsequent screening mammograms.

TABLE 1
Mammographic results according to initial or final conclusion, by round

	Ist round N = 12.239		Subsequent rounds N = 3.385	
	+	–	+	–
Initial conclusion	986 8,1%	11.253 91,8%	184 5,4%	3.201 94,6%
Final conclusion (after consensual double reading)	951 7,8%	11.288 92,2%	170 5,0%	3.215 95,0%

Mammographic results: Thousand hundred seventy mammograms were classified as abnormal by the first reader, so the "recall rate" at initial conclusion reached 7,5%. After the double reading process, 1.121 mammograms were classified as abnormal (recall rate at final conclusion of 7,2%). At first round, recall rate was 8,1% according to initial conclusion, and 7,8% according to final conclusion. Recall rate in subsequent round was 5,4% according to initial conclusion and 5,0% according to final conclusion (table 1).

Global recall rate within centres ranged between 3,7% and 11,0% at first reading; it was reduced to 4,8% through 10,7% after double reading.

Concordance: observed concordance (table 2) between readers was 93,7% (from the 15.624 mammograms, 13.880 negative and 778 positive results were concordant at both reading), while kappa coefficient was 58,7%. Positive agreement was 61,5%. and negative agreement was 96,6%.

TABLE 2
Concordance between readers

	Second reading		
	Negative	Positive	Total
First reading			
Negative	13.880	574	14.454
Positive	392	778	1.170
Total	14.272	1.352	15.624

Reclassification: globally, 267 (22,8%) of the 1.170 initially positive mammograms were reclassified as negative after the double reading process. 218 of the 14.454 (1,5%) of initially negative mammograms were reclassified as positive after the consensual double reading process.

Further assessment

1.388 mammograms classified positive at first or at final conclusion. We could obtain further assessment information for 1.249 of those 1.388 women (90%), while 139 could not be traced (no answer from the practitioner, change of practitioner).

Table 3a summarises the screening performance indicators (process and prognostic indicators), at first screening round. Of the 12.239 first screening mammograms, 951 (7,8%) were classified as "positive" for the screening test after consensual double reading; biopsy was performed in 1,1% of the screenees (1,0% open biopsy, 0,1% microbiopsy). Cancer detection rate reached 5,8‰, which correspond to a prevalence/expect-

TABLE 3a
Screening performance indicators at first screening round, compared with some "Europe against Cancer" norms

	1st round N = 12.239		"Europe against cancer" recommended norms	
			Acceptable	Desirable
Processus indicators				
Recall, according to final conclusion	951	*7,8%*	*< 7%*	*< 5%*
Open biopsy performed	120	*1,0%*	*< 0,5%*	*< 0,4%*
Microbiopsy (Thru-cut, stereotaxy) without open biopsy	13	*0,1%*		
Cancers detected	71	*5,8‰*		
DCIS/all cancers	15/71	*21%*		
Prevalence /expected incidence ratio		*2,9*	*3*	*> 3*
PPV of biopsy recommandation	71/133	*53,4%*		
benign/malignant ratio	62/71	*0,87:1*	*< 2:1*	*< 1:1*
Approximated specificity		*92,8%*		
Pronostic indicators				
Small size invasive cancer (< 10mm)		*40%*		*> 25%*
Invasive without nodal involvement		*65%*		

ed incidence ratio of 2,9 (17). Ductal in situ carcinoma (DCIS) represented 21% of all cancers (15/71). Positive predictive value of any biopsy indication was 53,4%, benign to malignant biopsy ratio was 0,87:1. The approximated specificity was 92,8%. At subsequent rounds (table 3b), recall rate was 5,0%, global biopsy rate was 0,7%, with an open biopsy rate of 0,6%. Cancer detection rate reached 5‰; DCIS proportion reached 53%; prevalence/expected incidence ratio was 2,5. Positive predictive value of any biopsy indication was 68%, benign to malignant biopsy ratio was 0,47:1, and approximated specificity reaches 95,5%.

Regarding prognostic indicators for detected cancers, 40% of invasive cancers detected at first round were smaller or equal to 10 mm, and 65% had no nodal involvement. At subsequent rounds, 33% of invasive cancers were smaller or equal to 10 mm, and 75% had no nodal involvement.

Additional imaging at the time of screening: additional ultrasound imaging was recorded since the 3rd year of the study. Echography was

TABLE 3b

Screening performance indicators at subsequent screening rounds, compared with some "Europe against cancer " recommended norms

	Subsequent rounds N = 3.385		"Europe against cancer" recommended norms	
			Acceptable	Desirable
Processus indicators				
Recall, according to final conclusion	170	*5,0%*	*< 5%*	*< 3%*
Open biopsy performed	20	*0,6%*	*< 0,35%*	*< 0,2%*
Microbiopsy (Thru-cut, stereotaxy) without open biopsy	5	*0,1%*		
Cancers detected	17	*5‰*	*1,5 x IR*	
DCIS/all cancers	9/17	*53%*		
Prevalence /expected incidence ratio		*2,5*		
PPV of biopsy recommandation	17/25	*68%*		
benign/malignant ratio	8/17	*0,47:1*	*< 1:1*	*< 0,5:1*
Approximated specificity		*95,5%*		
Pronostic indicators				
Small size invasive cancer (< 10mm)		*33%*		*> 25%*
Invasive without nodal involvement		*75%*		

registered at least in 21% of the patients at the first round screening, and in 24% of the patients at subsequent screening rounds. Indications and results of echography were not recorded.

Correlation between mammographic reading and histological findings: among the 158 biopsies performed at any round, 70 were benign and 88 were malignant (table 4). For 78 of the 88 malignant biopsies (89%) mammograms were classified as positive by both readers. Six cancers were seen only by the first reader and missed by the second (6,8%), and correctly reclassified at the third reading. Four were missed by the first reader and found by the second.

Improvement in detection rate was: (4 + 6/2) / (78 + (4 + 6)/2) = 6%.

Discussion

The first objective of this work was to evaluate the screening performance, by comparing observed outcomes with recommended Europeans norms and with outcomes from other projects. It can be argued that reference values admitted to judge the observed outcomes of a mass screening programme could not be suitable to assess a self-referred programme, because characteristics of women in those two types of programmes can be different. Self- referred women might have a greater motivation and health awareness than the general population and this could be related to a familial history or breast cancer, or a higher socio-economic or educational level, both factors associated with a

TABLE 4
Correlation between histological finding and mammographic conclusion

	Two readings positive	1st reading positive only*	2d reading positive only*	Total
Histological finding				
Benign	49	8	13	70
Malignant	78	6	4	88
	127	14	17	158

* Cases with discordant results at first two reading that were classified as positive at the consensual third reading.

higher risk of breast cancer. This self-selection bias would result in a higher prevalence of breast cancer within screened women. Because the positive predictive value of screening, and the benign to malignant biopsy ratio depend on prevalence as well as screening tests characteristics, good values for those indicators can be achieved because of a higher prevalence rather than good test performance. Therefore, good values for those two indicators must be interpreted with caution. At the contrary, poor results must certainly be considered. Anyway, this selection bias is also present, but in a fewer extend, in mass screening programmes: women with a greater health awareness are always more likely to attend. Because a pilot programme has the responsibility to test procedures before being extended, we nevertheless considered it was useful to proceed to an evaluation.

Overall recall rate was quite high, according to European norms, but was encouragingly decreasing between the 1st and the 4th year (data not shown). Recall rate in the centres ranged from 3,7% to 11,0%. Variability can partly be explained by self-selection bias that played in different directions from one centre to another. Radiologic centres have also heterogeneous screening throughput, resulting in some heterogeneity in radiologists experience. In Flanders, recall rate in different pilot projects ranged from 3,3% to 9,2% (7).

The "Europe against cancer" programme recommends that no more than 5% additional imaging be performed at the time of screening. In the Brussels project, the rate of additionnal echography reached at least 21%, which is uncommonly high. In Flanders, the only published echography rate (in Gent) was 2,3%. Our high rate can be explained by the fact that large scale, organised mammographic screening programme with precise guidelines have never been implemented in the French speaking part of Belgium (and only quite recently in the Flemish part). Quality assurance projects are working with existing facilities, which have developed their own screening protocols and habits. This should be taken into account in case of implementation of a large-scale screening programme: cumulating of dual reading and a high rate of immediate echography represents an unacceptably high financial burden.

Biopsy rate was higher than recommended by "Europe against Cancer", but remained within the range described in other projects: in France, the biopsy rate among 6 regional programmes ranged between 0,9 and 1,6% (18); in the UK National Health Service Breast Screening Programme, the biopsy rate was 1% (19).

Benign to malignant biopsy ratio was inferior than 1:1, which is a good value according to European norms. In the Flemish projects, benign to malignant biopsy ratio ranged from 0,5:1 to 1,2:1 (7).

Cancer detection rate was 2,9 times the natural incidence rate in the considered agegroup: this can be considered as acceptable according to European norms. Small cancer proportion and cancer without nodal involvement proportion were relatively high, indicating a high proportion of good prognosis cancers. This is an early proxy for efficacy of the screening process.

Currently, no assessment of the sensitivity has been done in any Belgian programme because the link with the National registry of cancer was not yet possible. The feasibility of this link has been studied and is now accepted by the ethical Committee authority of the Register and the Commission for the protection of privacy. Calculation of sensitivity should thus become possible in the next few years.

Our second objective was to evaluate the impact of double reading. Systematic double reading has been recommended by the programme "Europe against cancer" as part of a quality assurance programme. Though, models of double reading vary, involving either consensual decision on discordant interpretations (20), either recalling of all women positive at least at one reading(16;21). The level of experience of the second reader is also subject of debate, some projects involving experts radiologists as second reader (21), other equally trained radiologists. Brown (22) compares 3 strategies of reading applied to the same context of a screening programme involving a series of more than 33.000 women: a single reading, a consensual double reading and a non-consensual double reading. Compared with single reading, consensual double reading increased detection rate of 12,5%, and decreased the recall rate of 40%; non-consensual double reading increased detection rate of 13,5% but increased recall rate up to 40%.

In the Brussels Project, consensual double reading was chosen. This should have lower recall rate and improve detection rate; mean effect on detection rate was 6%, which is less marked than in other projects. In the Gent project (Belgium), mean gain in detection rate was 12,5% (15). Other studies achieved gain of sensitivity from 12 to 20% (16,21,23). Drop of recall rate was also slight, moving from 7,5% according to initial conclusion to 7,2% according to final conclusion. Maybe the high rate of immediate echography can explain the more limited impact of double

reading in our project; to test this hypothesis we should registrate indication and result of additional echographies. Impact of echography at time of screening and double reading could then been compared.

Interobserver variability has been studied in two ways: dedicated (or "adhoc") studies (24,25) and routine screening situations (15,20,21). Kappa coefficient is generally used to summarise interobserver variability; comparisons between studies should nevertheless be done with caution as kappa coefficient is influenced by the disease prevalence and the test sensitivity and specificity (14,20,26). Positive and negative agreement (14) should be considered besides kappa value. Kappa coefficient was relatively low (60%) in our study, but still in the range of most interobserver agreement in mammographic interpretation studies: in the regional French programme of the Bouche-du Rhone, a kappa coefficient of 67% was found. In Gent, Bleyen reported a kappa of 51%, in the New Zealand programme, a kappa value of 65% was reported. One dedicated study reproduced a simulated screening situation where 150 mammograms were read by 10 radiologists; mean kappa between pairs of radiologists was 47% (25). Another small dedicated study found a mean kappa of 48% (24). We found a positive agreement of 61%, which is relatively low, and a negative agreement of 95%. Main reasons for interobserver variability are difference of perception and difference in interpretation (27). Despite similar basic training and continual training process induced by dual reading, variability remains high. This is an argument supporting the need of ongoing with training of radiologists. Maybe the adjunction of a radiologist expert as second reader would improve results of the second reading.

In conclusion, the recall rate should further decrease. Additional imaging rate is unacceptably high in the perspective of an extension of the project. Other performance indicators in the Brussels Project compare quite well with other published results in similar context. Impact of double reading is low. This point should definitely be further explored.

Acknowledgements

The authors thank all radiologists implied in the project and all practitioners who helped them to collect data. They also gratefully thank Mrs Cannoodt for her supporting work.

References

1. SHAPIRO S, VENET W, STRAX P, VENET L, ROESER R. Ten to fourteen -year effect of screening on breast cancer mortality. J Natl Canc Inst 1982; 69: 349-355.
2. SHAPIRO S. Periodic screening for breast cancer: the HIP Randomized Controlled Trial.The Health Insurance Plan. J Natl Cancer Inst Monogr 1997; (22): 27-30.
3. TABAR L, GAD A, HOLMBERG L, DAY N. Reduction in mortality for breast cancer after mass screening with mammography. Lancet 1985; 1: 829-832.
4. TABAR L, FAGERBERG G, DUFFY SW, DAY NE. The Swedish two county trial of mammographic screening for breast cancer: recent results and calculation of benefit. J Epidemiol Community Health 1989; 43(2): 107-114.
5. DAY N, BAINES CJ, CHAMBERLAIN J, HAKAMA M, MILLER AB, PROROK P. UICC project on screening for cancer: report of the workshop on screening for breast cancer. Int J Cancer 1986; 38: 303-308.
6. Europe against Cancer. Recommendations of the Comittee of experts on Breast Cancer Screening. Copenhagen 1990. Annex 5; Brussels, European Community, DGV. 1990.
7. WEYLER J, SCHRIJVER J, VANDERMEEREN I, VUYLSTEKE DE LAPS L. Multicenterstudie bortskankerscreening Vlaanderen. 1998. Brussel, Ministerie van de Vlaamse Gemeenschap.
8. GORDENNE W. Preliminary results of a screening programme by mobile units in the province of Liège. J B R 1997; 80(3): 120-121.
9. DAY N, WILLIAMS D, KHAW K. Breast cancer screening programmes: the development of a monitoring and evaluation system. Br J Cancer 1989; 59: 954-958.
10. VERBEEK A, VAN DEN BAN M, HENDRIKS J. A proposal for short-term quality control in breast cancer screening. Br J Cancer 1991; 63: 261-264.
11. SICKLES E, OMINSKY S, SOLLITTO R, GALVIN H, MONTICCIOLO D. Medical audit of a rapd-throughput mammography screening practice: methodology and results of 27114 examinations. Radiology 1990; 175: 323-327.
12. Europe against Cancer. European guidelines for quality assurance in mammography screening. 2d edition. 1996. Brussels, Office for official publications of the European Communities.
13. DE LUYCK I, MOL H, VAN LOON R, OSTEAUX M. Technical quality control in mammography screening: first results in Belgium. J B R 1995; 78: 311-312.
14. CICCHETTI D, FEINSTEIN A. High agreement but low kappa:II. resolving the paradoxes. J Clin Epid 1990; 43(6): 551-558.
15. BLEYEN L. Apport de la lecture en double aveugle des mammographies dans la détection des cancers du sein. Medisphere 1997; 59: 37-40.
16. THURFJELL E, LERVENALL K, TAUBE A. Benefit of independent double reading in a population-based mammography screening programme. Radiology 1994; 191: 241-244.
17. HAELTERMAN M. Cancer en Belgique:1993-1994. 1998. Bruxelles, Registre National du Cancer.
18. WAIT S, ALLEMAND H. The French breast cancer screening programme: epidemiological and economic results of the 1st screening round. Eur J Public Health 1996; 6: 43-48.
19. PEETERS P, VERBEEK A, HENDRIKS J, VAN BON M. Screening for breast cancer in Nijmegen.Report of 6 screening rounds,1975-1986. Int J Cancer 1989; 43: 226-230.

20. WILLIAM S, DOYLE T, CHARTRES S, RICHARDSON A, ELWOOD J. Impact of independent double reading of mammograms from the inception of a population-based breast cancer screening programme. The Breast 1995; 4: 282-288.
21. SERADOUR B, WAIT S, DUBUC M. Dual reading in a non-specialised breast cancer screening programme. The Breast 1996; 5: 398-403.
22. BROWN J, BRYAN S, WARREN R. Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. B M J 1996; 312: 809-812.
23. Programme Mammographie du Luxembourg. Résultats de l'année 1996. Bulletin de liaison 1998.
24. CHICCONE G, VINEIS P, FRIGERIO A, SEGNAN N. Inter-observer and intra-observer variability of mammogram interpretation: a field study. Eur J Cancer 1992; 28A(6): 1054-1058.
25. ELMORE J, WELLS C, LEE C, HOWARD D, FEINSTEIN A. Variability in radiologist-s'interpretation of mammograms. N E J M 1994; 331: 1493-1499.
26. FEINSTEIN A, CICCHETTI D. High agreement but low kappa:1.the problems of two paradoxes. J Clin Epid 1989; 43: 543-549.
27. GOERGEN S, EVANS J, COHEN G, MACMILLAN J. Characteritics of breast carcinomas missed by screening radiologists. Radiology 1997; 204: 131-135.