A Novel Data Analytics-derived Metric (Nearest Cluster Distance) Is Easily Implemented in Routine Practice and Correctly Identifies Breast Cancer Cases for Quality Review

Richard E. Whisnant *

HCA Florida Healthcare, Fawcett Memorial Hospital, Department of Pathology, 21298 Olean Blvd, Port Charlotte, FL 33952, USA

ARTICLE INFO

ABSTRACT

Background: Errors in breast cancer grading and predictive testing are clinically important and can be difficult to detect in routine practice. A quality metric able to identify a subset of breast cancer cases which are high yield on quality review would be of practical clinical benefit.

Methods: Data analytic techniques were used to generate consensus tumor signature centers from a dataset over 500 breast cancer cases from a single practice. Cases were assigned a novel metric, Nearest Cluster Distance, corresponding to their distances from the nearest tumor signature center. The subset of tumors exceeding a cutoff for this metric were flagged, and then reviewed and rescored in a blinded fashion together with matched controls. A simplified version of this metric was created using universally accessible methods.

Results: Flagged cases showed statistically significant movement toward consensus tumor signature centers compared with controls, consistent with identification of cases which could benefit from review and possible rescoring. The simplified metric performs identically.

Conclusion: This method can be readily applied in routine practice and is promising as a real time quality check for breast cancer diagnosis and reporting.

Background

Breast cancer treatment and risk assessment are critically dependent on accurate diagnosis and reproducible assessment of histologic parameters (grade) and results of predictive testing (ER/PR/Her2Neu). These multiple parameters are not independent, but rather constrained by certain allowed combinations, corresponding to the known subgroups of breast cancer. Though these groups were defined in part by cluster analysis of gene expression data,[1] subsequent work has shown that unsupervised clustering of histologic and immunohistochemical data can generate clinically valid groupings.[2,3] In this approach, the combination of seven parameters (three grade elements, lobular vs. ductal differentiation, ER, PR, and Her2 status) locates each breast cancer case within seven-dimensional space, and the known subgroups (such as luminal-A or triple-negative breast cancer) exist as clusters of cases occupying constrained regions in this space. Cases located outside of these clusters should be rare and viewed as suspect, indicative of either an error in one or more parameters, or a highly unusual tumor.

Historically, high levels of such errors were reflected in frequent discordance between labs[4,5] and widely publicized population-level failures.[6] These prompted sweeping changes in predictive testing, including proficiency testing and standardization in testing and reporting, with marked improvement in quality and patient care.[7–11] In spite of these improvements, challenges remain, especially at the level of individual, rather than systemic, errors in testing. In a recent study of ER + /Her2- breast cancers originally tested in the UK from 2012 to 2014, Pinder et al found a high (>95%) concordance of results compared with central review.[12] However, even after excluding all cases where explanations other than outright error could be found (such as tumor heterogeneity or borderline/low level expression/amplification), 1.6% of reviewed tumors showed major discrepancies. A similar population of purported ER + patients reclassified as ER- on central review showed poor disease-free survival in a separate study.[13] For perspective, this 1.6% rate applied to projected new breast cancer diagnoses would correspond to more than 4500 US patients in 2021 with major predictive testing errors.[14]

Simple quality metrics have been widely adopted which leverage the clustered nature of correctly assessed grade and predictive testing results. Examples include ASCO/CAP recommendations for repeat testing in certain scenarios (low histologic grade with weak or absent estrogen receptor expression, and Her2 overexpression with low nuclear grade) and CAP accreditation standards requiring surveillance for an appropriately low proportion of "low-positive" estrogen expressing tumors.[15] These guidelines

---

* Corresponding author.
  E-mail address: Richard.Whisnant@hcahealthcare.com.

essentially propose reconsideration of cases which are located in suspect regions of space; however, such guidelines generally include a limited number of parameters, because of the difficulty in visualizing and implementing more complex rules involving higher dimensional space. Looking for errors in certain regions of space because they are easy to visualize or evaluate is analogous to looking for lost items under a lamppost because that is where the light is.

A quality metric that readily identifies additional cases that exist outside of the "appropriate" regions of seven-dimensional might prove useful in identifying or preventing errors. The author hypothesized that a breast cancer dataset could be scaled and clustered in a manner which might provide the basis for identifying clinically important errors. A simple metric expressing the distance of any given case from the nearest of those cluster centers would highlight cases located in suspect regions of space. Such cases might benefit from reexamination or quality review.

Methods

Case Population and Construction of Dataset

All breast cancer specimens from a single practice (Suncoast Pathology; Venice, FL) from August 2016 through May 2020 for which results for ER, PR, Her2, and histologic grading were available on the same specimen were retrieved (548 cases). Tumors were graded throughout this period according to the Nottingham method.[16] Initial evaluation of ER/PR/Her2 status was performed by immunohistochemistry (IHC) [Neogenomics (Fort Myers, FL and Aliso Viejo, CA), ER clone 6F11, PR clone PgR1294, and Her2 clone 4B5]. All stains were interpreted by experienced pathologists in this hospital-based community practice, scored digitally (Indica Labs and Definiens platforms, Neogenomics), and reported according to ASCO-CAP guidelines. Her2 in-situ hybridization was performed on IHC $2+$, grade 1 Her2 $3+$, and grade 3 Her2 $0-1+$ cases.

Identifying case numbers were removed and the cases given unique identifiers. The correlation of unique identifiers and original case numbers was maintained by an honest broker and was inaccessible to the study pathologist. The study design was approved by the Institutional Review Board of Lutheran Hospital, Fort Wayne, IN. The Nottingham scores were coded as individual nuclear, tubule, and mitotic rate scores 1-3 each. ER and PR expression were expressed on a 0-300 scale (% of cells positive times strength). Her2 status was coded as positive for cases with $3+$ IHC scores or $2+$ IHC scores with amplified FISH results, and negative for all others. Lobular carcinoma status was coded as positive for conventional lobular carcinoma, and negative for both pleomorphic lobular and mixed ductal and lobular differentiation tumors.

Scaling and Clustering of Dataset

The data were then scaled as follows: Nottingham grade components each divided by 3, ER divided by 100, PR divided by 200, HER2+ as 3, and lobular carcinoma as 2. These scalings were chosen to conform to the perceived likelihood and importance of potential errors. ER and Her2 were felt to be much more often critical than PR or lobular differentiation status, and so received heavier weightings. Combined histologic grade (rather than each of the three components) was set equal in scaling and importance to ER or Her2 status. Similarly, Her2 was made discrete rather than continuous based on the judgement that $2+$ Her2 results would be overwhelmingly accurately classified by reflex in-situ testing, and therefore were unlikely to represent clinical false-positive or false-negative results. This scaled dataset consisting of 548 cases was then subjected to unsupervised k-means clustering with 8 cluster centers [R version 4.0.1, The R Foundation for Statistical Computing, *kmeans* function with 100 initial configurations sampled (nstart=100)]. The number of cluster centers chosen was the smallest number of cluster centers which resulted in clusters which were uniform for Her2 and lobular status. This was implemented so that the clusters were "human readable" in the sense that they made

sense to a practicing pathologist and roughly corresponded to known breast cancer subtypes.

Post-clustering Calculation of Nearest Cluster Distances-Euclidean (NCD-E's)

The Her2+ lobular group was removed due to its small size. A significant number of these cases were familiar to the study pathologist, as they were from a single patient with multiple biopsies of multifocal disease with testing requested on all.

For each case, its Euclidean distance to each of the seven cluster centers was calculated using the seven parameter values of the case and the consensus center by the formula $\left(\sum_{n=1}^{7} (Case\_value_n - Consensus\_value_n)^2 \right)^{1/2}$. The minimum of these values was recorded as the case's nearest cluster distance-Euclidean (NCD-E) and the case assigned to that cluster. The distribution of distances was visually inspected, and a cutoff of ≥ 0.7 was designated for selection of cases for further analysis.

Contributors to Distance from Center

For the cases with NCD-E's ≥ 0.7, the Manhattan distance to the designated cluster center (nearest cluster distance-Manhattan (NCD-M)) was calculated by the formula $\sum_{n=1}^{7} |Case\_value_n - Consensus\_value_n|$. For each case, the proportion of this distance accounted for by grading parameters (grade_proportion) was calculated [(distance_from_nuclear_grade_center + distance_from_tubule_formation_center + distance_from_mitotic_rate_center)/ total_manhattan_distance]. As all lobular+ and Her2+ cases clustered together and these parameters are binary, the contribution of the hormone receptor scores (receptor_proportion) is equivalent to (1 - grade_proportion).

Blinded Rescoring

For each flagged case, a matching control case was selected at random from cases assigned to the same cluster. All cases where all diagnostic slides were available in our files (32 flagged, 33 controls) were retrieved by the honest broker and rescored by the study pathologist blinded to both original diagnosis and identity of cases as flagged or control cases. Grading was done as originally. Rescoring of ER/PR/Her2 was manual. Hormone receptors were scored by visual estimation of total percent of cells positive and predominant strength $(1–3+)$ among those cells, with the hormone receptor level the product of those numbers (0–300). Her2 was scored preliminarily as definite negative $(0–1+)$ or possible positive $(2–3+)$. Possible positives were assigned a final rescoring classification based on prior computer assisted scoring or FISH testing results, accessed by the study honest broker.

These rescored results were then used to calculate post-rescoring NCD-E values. Movement of cases toward nearest cluster center was calculated as (NCD-E$_{original}$–NCD-E$_{rescored}$), resulting in a positive number if a case is closer to the nearest cluster center after rescoring and negative if further away. Movement was compared for flagged and control populations with the Wilcoxon rank sum test.

Algorithm for Determining Modified Cluster Assignments and Calculating Modified Cluster Center Locations

The dataset was divided randomly into equal-sized training and test sets. An algorithm using cutoffs for ER, PR, Her2 status, and lobular status was developed by trial and error on the training set, with the goal of maximizing the number of cases classified identically to the original cluster classification. This algorithm was then applied to the test set and these results also compared to the original classifications. For each modified cluster of cases defined as above, the mean was calculated for each of the seven parameters and this combination of values designated the modified cluster center location for that cluster. The Euclidean distance from the original cluster center to the modified one was calculated for each cluster as

Table 1

The eight calculated clusters, their consensus parameter values, and the proportion of cases assigned to each. Note: cluster center values rescaled for readability.

Cluster name/ Tumor type	Number of cases	Percent of cases	Nuclear score	Tubule score	Mitotic rate score	ER level	PR level	Her2 status (Pos = 3)	Lobular status (Pos = 2)
Luminal A	195	36%	2.0	2.3	1.1	295	268	0.0	0.0
Luminal B	130	24%	2.0	2.3	1.2	289	34	0.0	0.0
Lobular "A"	41	7%	1.5	3.0	1.0	286	234	0.0	2.0
Lobular "B"	45	8%	1.5	3.0	1.0	279	28	0.0	2.0
Her2 +	33	6%	2.9	2.9	1.6	8	0	3.0	0.0
Luminal Her2 +	42	8%	2.5	2.8	1.2	286	82	3.0	0.0
TNBC	54	10%	2.9	2.8	2.0	11	0	0.0	0.0
Her2 + lobular	8	1%	1.1	2.9	1.0	267	124	3.0	2.0

above. This entire process was performed in Excel, and a template spreadsheet with all required steps and calculation fields is available on request.

Calculation of Modified NCD-E's

The NCD-E$_{modified}$ was calculated in Excel for each case in the entire dataset using the modified cluster centers. The sensitivity and specificity of an NCD-E$_{modified}$ cutoff of ≥ 0.7 for identifying the previously flagged cases was determined.

Results

Characterization and Distribution of Clusters Assigned

This clustering method yielded clusters which classified all Her2 + and all lobular cases together as designed. The cluster center locations and their frequencies within this dataset are summarized in Table 1. For the original diagnoses, scaling and clustering appears appropriate, with the large majority of the variance within the final dataset accounted for by between cluster distance as opposed to within cluster distance (total between cluster sum of squares = 1649 out of total sum of squares = 1760).

The tumor type labels in Column 1 are descriptive labels for purposes of discussion - some correspond to traditional classification of breast cancers (i.e., Her2 + and TNBC), while others do not. Note that this clustering method assigns cluster centers in a way that is outlier sensitive, so the cluster centers are shifted slightly from the levels shown by the large majority of cases within that cluster. For example, the TNBC cluster shows low level ER expression, though the large majority of cases in this cluster are completely negative for ER. These outliers were not removed, as no trivial method for k-means clustering that de-emphasizes outliers is available.
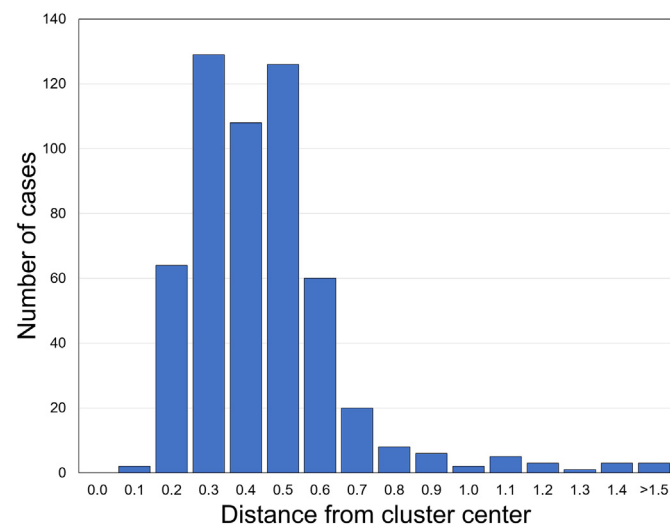
Distribution of NCD-E's

The distribution of NCD-E's for the complete dataset is shown in Fig. 1. The large majority of cases (94%) are tightly clustered around the median, consistent with highly accurate grading and scoring of these cases. The distribution shows a shoulder and long positive tail, with 35 cases having NCD-E's greater than or equal to 0.7, and 23 of these ranging from 0.83 to 2.0. Each of the seven clusters is represented by at least 1 case among these 35.

The distribution is bimodal and shifted from zero, as are the distributions for cases within the individual clusters (data not shown). This is expected, based on the binary nature of the parameters, combined with the presence within each cluster of multiple parameter value cases. That is, no case has a tubule score of 2.3, and therefore all cases will be a non-zero distance from that cluster center.

Parameter Contributions to NCD

The Manhattan distances to cluster centers (NCD-M's) were calculated for each of these cases. Manhattan distances will be greater than or equal to Euclidean distances, by an amount based on the number of different parameters contributing to distance. The 35 cases selected show a range of NCD-E's of 0.7-2.0 and the corresponding NCD-M's range from 1.0 to 2.6.

The NCD-M's are the sum of the contributions from each of the five components: each of the three Nottingham scores, and the two hormone receptor scores. Each of these five was the main contributor to distance in at least one case. Because all clusters are uniform for Her2 and lobular status (all positive or negative for each), neither of these parameters can contribute to NCD. The distribution of proportion of distance contributed by grade

Figure 1. Distribution of NCD-E's, complete dataset. Mean = 0.40, median = 0.38, SD = 0.21, and IQR = 0.21
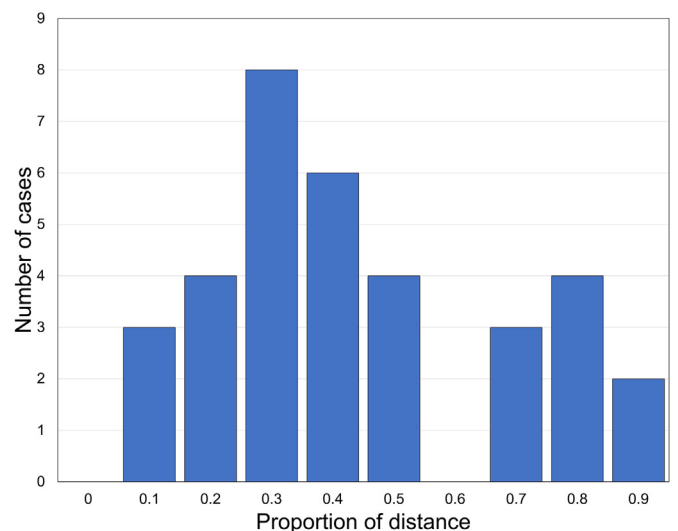
Figure 2. Proportion of NCD-M accounted for by grade parameters (1 – hormone receptor proportion), 35 flagged cases.

parameters is shown in Fig. 2. Grade was the predominant contributor in 9 (of 35) cases; hormone receptor scores were the main contributor in the remaining 26 cases.

Results of Rescoring

The distributions of movement toward nearest cluster on rescoring for test and control groups is shown in Fig. 3. Test cases showed a significantly higher median movement toward nearest cluster ($p < 10^{-5}$ by Wilcoxon rank sum test). Control cases showed median and mean movement of close to zero (–0.03 and –0.08, respectively).

Modified Cluster Assignments

The manually generated algorithm criteria for modified cluster assignments are shown in Table 2. In the training set, this algorithm assigned 268 of 270 cases to the same cluster as the original analysis. In the test set, 269 of 270 cases received the same assignment as originally. All three

Table 2
Modified cluster centers summary: inclusion criteria and distance of modified cluster centers from original data analytic derived centers.

Cluster name/ Tumor type	Inclusion criteria				Distance to original cluster
	ER level	PR level	Her2 status	Lobular status	
Luminal A	>100	≥160	Negative	Negative	0.03
Luminal B	>100	<160	Negative	Negative	0.03
Lobular "A"	>100	>140	Negative	Positive	0.01
Lobular "B"	>100	≤140	Negative	Positive	0.05
Her2 +	<100	NA	Positive	Negative	0.03
Luminal Her2 +	>100	NA	Positive	Negative	0.00
TNBC	<100	<100	Negative	Negative	0.03

cases which failed to classify as originally showed original NCD-E's greater than 1. The distances from original to modified cluster centers is also shown in Table 2.

Using the cutoff of ≥0.7, the identical 35 cases flagged with NCD-E$_{modified}$ as had been previously identified using NCD-E$_{original}$. The correlation of NCD-E$_{modified}$ and NCD-E$_{original}$ for these cases is shown in Fig. 4.

Discussion

Nearest Cluster Distance (NCD) shows promise as a quality metric for breast cancer diagnosis and prognostic testing. Unsupervised clustering of a scaled dataset populated by breast cancer cases in a community practice generated cluster centers consistent with the biology of breast cancer subtypes. Most cases were located close to one of these clusters, with review showing very little change from initial parameters. In contrast, rescoring of the small subset (6%) of cases located in the shoulder or long tail of the NCD distribution resulted in significantly lower NCD's – i.e., moved them closer to a cluster center on review. The decrease in NCD was highly significant compared to controls. This outcome strongly suggests that implementing a metric of this kind has the potential to flag cases for review in real time, potentially preventing misclassifications.

This study also validates a modified version of this metric which can be implemented in virtually any practice setting. While NCD was developed with data analytic tools, specifically k-means clustering, these tools are not in broad use. Therefore, the modified approach was developed to generate the centers needed in virtually any spreadsheet application. With practice specific cluster centers in hand, this approach is trivial to implement in routine practice. Figure 5 shows a simple
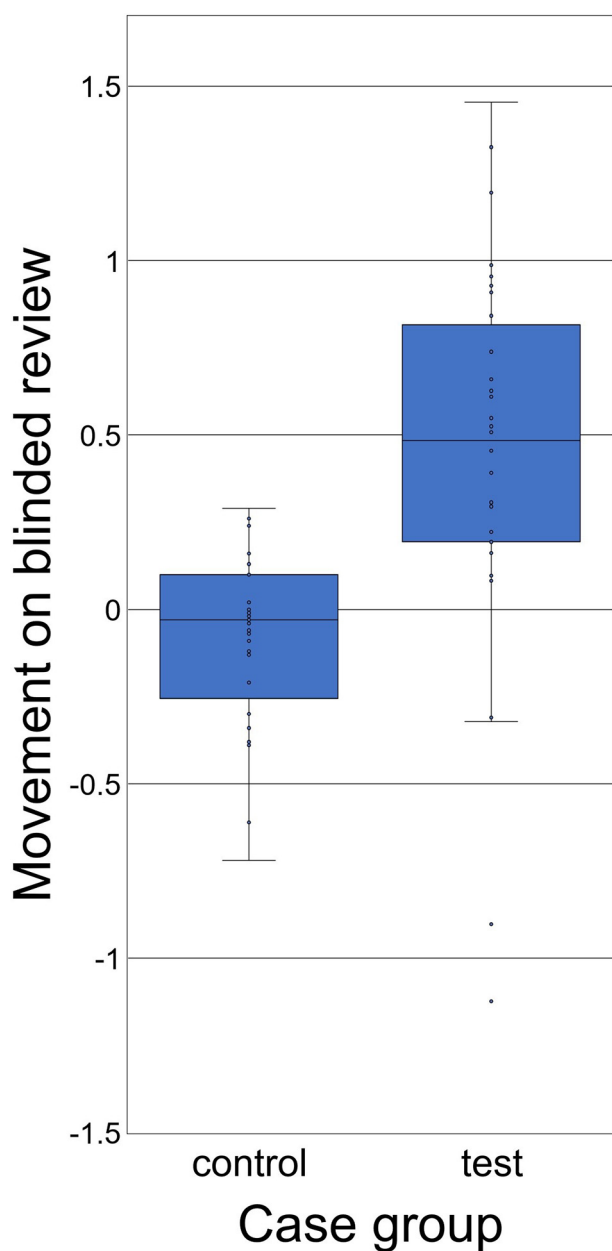

Figure 3. Movement of cases toward nearest cluster on blinded review.
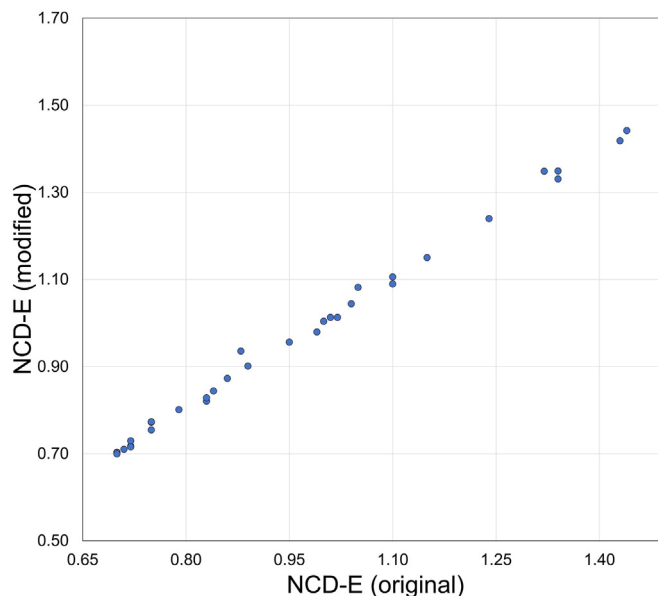

Figure 4. Correlation of NCD-E's (original versus modified) for the 35 flagged cases.

Enter your consensus cluster centers here:

Cluster name	Nuc grade	Tub form	Mit rate	ER	PR	Her2	lob
Luminal A	2	2.3	1.1	295	268	0	0
Luminal B	2	2.3	1.2	289	34	0	0
Lobular "A"	1.5	3	1	286	234	0	2
Lobular "B"	1.5	3	1	279	28	0	2
Her2+	2.9	2.9	1.6	8	0	3	0
Lum Her2+	2.5	2.8	1.2	286	82	3	0
TNBC	2.9	2.8	2	11	0	0	0

Enter your cases to be assessed here: | Quality metric | Spreadsheet calculates distances here:

Distances of case to cluster centers:

Case #	Nuc grade	Tub form	Mit rate	ER	PR	Her2	lob	NCD-E	lum A	lum B	lob A	lob B	Her2+	lum Her2+	TNBC
S-22-1234	2	2	1	286	67	0	0	0.21	1.02	0.21	2.20	2.04	4.13	3.02	2.81
S-22-1235	3	3	3	15	0	0	0	0.34	3.19	2.84	3.66	3.42	3.04	4.11	0.34
S-22-1236	2	2	1	180	40	0	0	1.10	1.62	1.10	2.49	2.26	3.50	3.20	1.78
S-22-1237	1	1	1	297	286	0	0	0.56	0.56	1.38	2.13	2.48	4.50	3.27	3.33
S-22-1238	3	3	1	4	0	3	0	0.21	4.41	4.16	4.75	4.56	0.21	2.85	3.02
S-22-1239	3	3	1	2	0	3	0	0.22	4.42	4.18	4.77	4.58	0.22	2.88	3.02
S-22-1234	2	3	1	298	294	0	0	0.27	0.27	1.33	2.03	2.42	4.44	3.19	3.26
S-22-1240	3	3	2	0	0	0	0	0.13	3.28	2.93	3.73	3.49	3.00	4.18	0.13
S-22-1241	3	3	2	200	300	3	0	1.43	3.19	3.43	3.77	3.98	2.44	1.43	3.85
next case here															

Figure 5. Screenshot of a screening tool for implementing this approach. Excel file available on request.

spreadsheet interface for screening single or multiple cases. Entry of the seven parameters in the case list automatically calculates the NCD-E and highlights any values of the metric above a selected cutoff. This process could also be automated, e.g. within APLIS systems, with real-time decision-support provided on parameters extracted from structured data such a cancer reporting protocols.

Each of the seven tumor types showed at least one flagged case, and the distances of these cases from nearest center were distributed broadly across all three histologic parameters and both hormone receptors. This is a particular strength of this approach - that it has the ability to identify many different types of errors which may be present in a case or series of cases. It should be noted that not all potential errors are highlighted equally with this approach; an isolated error that moves a case near to an incorrect cluster center will be invisible to this metric.

There are strengths and weaknesses of this study. The number of cases included is large enough to support successful clustering, though Her2 + cases (14%) and TNBC cases (10%) are relatively underrepresented due to the patient population served in our practice. Diagnosis and predictive testing were uniformly performed by a small number of pathologists (three), and second review of diagnosis and grading was done for the large majority of the cases. The primary weakness of this study is its retrospective blinded design. While the findings strongly suggest that the subpopulation identified by this approach may have benefited from reexamination, the exact changes from original to review diagnosis cannot be determined for any individual case. Therefore, the clinical importance of these changes cannot be assessed. Assessment of a non-blinded dataset might allow refinement of this metric to maximize sensitivity and specificity for the most clinically important findings. In addition, because only previously generated predictive testing slides were reviewed, this study only evaluates post-analytic (interpretive) performance.

The utility of this approach needs to be validated in independent datasets. A non-blinded study would allow recalculation of the cluster centers by excluding data points which on review were judged to be inaccurate. This adjustment might serve to more reliably highlight meaningful outliers. In addition, refinement of the metric itself, including exploration of different distance metrics, might improve sensitivity and specificity. NCD as implemented in the present study might prove useful in a wide variety of clinical practice settings.

Competing interests

The author declares that he has no competing interests.

Acknowledgments

References

1. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415(6871):530–536.
2. Soria D, Garibaldi JM, Ambrogi F, et al. A methodology to identify consensus classes from clustering algorithms applied to immunohistochemical data from breast cancer patients. Comput Biol Med 2010;40(3):318–330.
3. Soria D, Garibaldi JM, Green AR, et al. A quantifier-based fuzzy classification system for breast cancer patients. Artif Intell Med 2013;58(3):175–184.
4. Roche PC, Suman VJ, Jenkins RB, et al. Concordance between local and central laboratory HER2 testing in the breast intergroup trial N9831. J Natl Cancer Inst 2002;94(11): 855–857.
5. Perez EA, Suman VJ, Davidson NE, et al. HER2 testing by local, central, and reference laboratories in specimens from the North Central Cancer Treatment Group N9831 intergroup adjuvant trial. J Clin Oncol 2006;24(19):3032–3038.

6. The Commission of Inquiry on Hormone Receptor Testing, the Honourable Margaret A. Cameron, Commissioner. Commission of Inquiry on Hormone Receptor Testing, Volume 1: Investigation and Findings. https://www.cihrt.nl.ca/Final%20Report/volume%201%20Investigation%20and%20Findings.pdf March 2009. Accessed July 12, 2021.

7. Allison KH, Hammond ME, Dowsett M, et al. Estrogen and progesterone receptor testing in breast cancer American Society of Clinical Oncology/College of American Pathologists Guideline update. Arch Pathol Lab Med 2020;144:545–563.

8. Hammond ME, Hayes D, Dowsett M, et al. American Society of Clinical Oncology/College of American Pathologists Guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. Arch Pathol Lab Med 2010;134:907–922.

9. Wolff AC, Hammond ME, Allison KH, et al. Human epidermal growth factor receptor 2 testing in breast cancer American Society of Clinical Oncology/College of American Pathologists clinical practice guideline focused update. Arch Pathol Lab Med 2018;142:1364–1382.

10. Wolff AC, Hammond ME, Hicks DG, et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. J Clin Oncol 2013;31:3997–4013.

11. Wolff AC, Hammond ME, Schwartz JN, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. Arch Pathol Lab Med 2007;131:18–43.

12. Pinder SE, Campbell AF, Bartlett JMS, et al. Discrepancies in central review re-testing of patients with ER-positive and HER2-negative breast cancer in the OPTIMA prelim randomised clinical trial. Br J Cancer 2017;116:859–863.

13. Viale G, Regan MM, Maiorano E, et al. Prognostic and predictive value of centrally reviewed expression of estrogen and progesterone receptors in a randomized trial comparing letrozole and tamoxifen adjuvant therapy for postmenopausal early breast cancer: BIG 1-98. J Clin Oncol 2007;25(25):3846–3852.

14. American Cancer Society. How Common Is Breast Cancer? https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html#:Revised May 2021 Accessed July 3, 2021.

15. Anatomic Pathology Checklist. *ANP.22970 Annual Result Comparison - Breast Carcinoma. College of American Pathologists. Revised 6/4.* 2020.

16. Protocol. for the Examination of Biopsy Specimens from Patients With Invasive Carcinoma of the Breast. cap.org https://documents.cap.org/protocols/cp-breast-invasive-biopsy-20-1100.pdf. Published February 2020. Accessed 4/22/2021.