Deep learning assessment of breast terminal duct lobular unit involution: Towards automated prediction of breast cancer risk

Suzanne C. Wetstein  
Eindhoven University of Technology  
Ying Liu  
Washington University School of Medicine in St. Louis  
et al.

Follow this and additional works at: https://digitalcommons.wustl.edu/open_access_pubs

Please let us know how this document benefits you.

Recommended Citation

https://digitalcommons.wustl.edu/open_access_pubs/9333

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker.  
For more information, please contact vanam@wustl.edu.
Deep learning assessment of breast terminal duct lobular unit involution: Towards automated prediction of breast cancer risk

Suzanne C. Wetstein1*, Allison M. Onken2, Christina Luffman3, Gabrielle M. Baker2, Michael E. Pyle2, Kevin H. Kensler2, Ying Liu4, Bart Bakker5, Ruud Vlutters6, Marinus B. van Leeuwen5, Laura C. Collins3, Stuart J. Schnitt6, Josien P. W. Pluim1, Rulla M. Tamimi7, Yujing J. Heng8†, Mitko Veta9

1 Medical Image Analysis Group, Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands, 2 Department of Pathology, Harvard Medical School, Beth Israel Deaconess Medical Center, Boston, Massachusetts, United States of America, 3 Division of Population Sciences, Dana Farber Cancer Institute, Boston, Massachusetts, United States of America, 4 Division of Public Health Sciences, Department of Surgery, Washington University School of Medicine and Alvin J. Siteman Cancer Center, St Louis, Missouri, United States of America, 5 Philips Research Europe, High Tech Campus, Eindhoven, The Netherlands, 6 Dana-Farber/Brigham and Women’s Cancer Center, Harvard Medical School, Dana-Farber Cancer Institute-Brigham and Women’s Hospital, Boston, Massachusetts, United States of America, 7 Channing Division of Network Medicine, Department of Medicine, Harvard Medical School, Brigham and Women’s Hospital, Boston, Massachusetts, United States of America

† These authors are co-senior authors on this work.
* s.c.wetstein@tue.nl

Abstract

Terminal duct lobular unit (TDLU) involution is the regression of milk-producing structures in the breast. Women with less TDLU involution are more likely to develop breast cancer. A major bottleneck in studying TDLU involution in large cohort studies is the need for labor-intensive manual assessment of TDLUs. We developed a computational pathology solution to automatically capture TDLU involution measures. Whole slide images (WSIs) of benign breast biopsies were obtained from the Nurses’ Health Study. A set of 92 WSIs was annotated for acini, TDLUs and adipose tissue to train deep convolutional neural network (CNN) models for detection of acini, and segmentation of TDLUs and adipose tissue. These networks were integrated into a single computational method to capture TDLU involution measures including number of TDLUs per tissue area, median TDLU span and median number of acini per TDLU. We validated our method on 40 additional WSIs by comparing with manually acquired measures. Our CNN models detected acini with an F1 score of 0.73±0.07, and segmented TDLUs and adipose tissue with Dice scores of 0.84±0.13 and 0.87±0.04, respectively. The inter-observer ICC scores for manual assessments on 40 WSIs of number of TDLUs per tissue area, median TDLU span, and median acini count per TDLU were 0.71, 0.81 and 0.73, respectively. Intra-observer reliability was evaluated on 10/40 WSIs with ICC scores of >0.8. Inter-observer ICC scores between automated results and the mean of the two observers were: 0.80 for number of TDLUs per tissue area, 0.57 for median TDLU span, and 0.80 for median acini count per TDLU. TDLU involution measures evaluated by manual and automated assessment were inversely associated with age and menopausal status.

We developed a computational pathology method to measure TDLU involution. This
technology eliminates the labor-intensiveness and subjectivity of manual TDLU assessment, and can be applied to future breast cancer risk studies.

Background

Most benign breast lesions and breast cancers arise in the terminal duct lobular units (TDLUs) [1], the milk-producing structures of the breast. Russo et al. [2] historically classified TDLUs into four lobule types: type 1 (least developed; <12 acini/lobule), type 2 (evolves from type 1; intermediate in degree of differentiation; between 12 and 80 acini/lobule), type 3 (fully developed structures; >80 acini/lobule), and type 4 (occurs during pregnancy and lactation).

Pathologists have used these qualitative lobule types to evaluate TDLU involution indicated by the presence of more type 1 lobules and less type 2 and 3 lobules after the completion of child-bearing and with physiological aging [3]. In quantitative terms, TDLU involution is characterized by a reduction of the size of TDLUs, the number of acini, and the number of acini per TDLU [4–8]. Previous work by our group and others evaluated TDLU involution using qualitative measures and reported that women with less TDLU involution (i.e., majority of lobules were of types 2 and 3) were more likely to develop breast cancer compared to those with predominantly type 1 lobules independent of age [5, 9, 10, 11]. Thus, TDLU involution measures may be utilized as a biomarker to assess breast cancer risk [9, 10].

Efforts to develop quantitative measures of TDLU involution started with McKian et al. [11] who evaluated the number of acini and TDLU area on histopathological sections. Rosebrock et al. [12] were the first to automatically estimate quantitative measurements from TDLUs and use those measurements to describe and classify them. Later, Figueroa et al. standardized three quantitative measures of TDLU involution—number of TDLUs per tissue area (TDLUs/mm$^2$), median TDLU span, and the median number of acini per TDLU (median acini/TDLU)—by assessing up to 10 TDLUs in the normal tissue for a WSI [4, 10, 13, 14]. The examined tissue area was corrected for the amount of adipose tissue present. These quantitative measurements still relied on manual histological assessment of breast tissue, and remained subjective and labor-intensive. Thus, the need for manual qualitative and/or quantitative assessment by pathologists is a major bottleneck to studying TDLU involution in large epidemiological studies.

Automated image analysis methods have the potential to decrease the workload of pathologists and standardize clinical practice [15]. Known or novel tissue biomarkers can now be automatically quantified [15–20] and deep learning has also been applied to recognize morphological tissue patterns for diagnostic purposes [21–27]. More specifically, networks have been successfully developed for tasks in breast histopathology [28–33]. Most recently, state-of-the-art deep convolutional neural networks (CNN) have been shown to outperform pathologists in detecting metastases in sentinel lymph nodes of breast cancer patients [34]. In this study, we developed an automated method to quantitatively assess TDLU involution. First, we constructed and optimized three deep neural networks to detect and/or segment acini, TDLUs, and adipose tissue. These three networks were integrated into a single method to compute TDLU involution measures. Our automated method was validated by comparing the automated measures with manually acquired measures on an independent set of images.

Methods

Subjects and acquisition of images

The participants in this study are from the Nurses’ Health Study (NHS) and NHSII. The NHS was established in 1976 with 121,700 US female registered nurses between 30–55 years of age,
and NHSII was established in 1989 (n = 116,429, ages 25–42). All NHS/NHSII participants are followed up biennially to obtain updated information on a range of epidemiological data and identify newly diagnosed diseases [35]. Hematoxylin and eosin (H&E) breast tissue slides were retrieved for women who reported a biopsy-confirmed benign breast disease (BBD) and gave permission to review their biopsy records and original H&E slides [36–42]. The tissue was prepared and stained at the local centers and centrally reviewed. BBD H&E whole slide images (WSIs) were obtained by scanning the slides at ×40 magnification with a resolution of 0.16 μm per pixel using Pannoramic SCAN 150 (3DHISTECH Ltd, Budapest, Hungary). The study protocol was approved by the institutional review boards of the Brigham and Women’s Hospital and Harvard T.H. Chan School of Public Health, and those of participating registries as required. Informed consent was obtained from all NHS/NHSII participants.

Developing the automated method for TDLU involution measures

In total, 92 WSIs from 92 benign breast biopsies from 67 pre- and 25 post-menopausal women were randomly selected from the NHS database. To capture the large variability in lobule sizes, pre-menopausal women were over selected to obtain data/annotation/ground truth for type 2 and 3 lobules since post-menopausal tissues were predominantly type 1 lobules. Due to the more challenging nature of the TDLU segmentation task, 92 WSIs were used to develop the TDLU segmentation neural network model while a subset of 50 out of the 92 WSIs was adequate to develop the acini detection and adipose tissue segmentation neural network models. Breast tissue with more adipose tissue has fewer TDLUs and acini [4], which influences the outcomes of TDLU involution measures (e.g. number of TDLUs per tissue area). Therefore, the adipose tissue model was developed to estimate and account for the percentage of adipose tissue.

TDLUs, acini, and adipose tissue were annotated within a region of interest (ROI) comprising approximately 10%, 10%, and 2.5% of the total tissue area, respectively. Annotation was done using the open-source software Automated Slide Analysis Platform (ASAP; Computation Pathology Group, Radboud University Medical Center). TDLUs were defined as clusters of acini in a lobular configuration. TDLU boundary was defined by the non-specialized/extra-lobular stroma. In order to assess involution in histologically normal breast parenchyma only, TDLUs with proliferative or metaplastic changes were not annotated as TDLUs but remained as background. Acini were defined as small spherical structures lined by epithelial cells and surrounded by myoepithelial cells. Acini with elongated shapes, epithelial proliferation, apocrine metaplasia, or without lumina were not annotated. In total, 25,645 acini and 1,631 TDLUs were annotated. Fig 1 shows examples of annotated acini, TDLUs and adipose tissue.

Acini, TDLUs, and adipose tissue were detected and segmented using the U-Net CNN architecture [43, 44]. Since we had different datasets for the three tasks, three separate models were trained. To construct the acini and adipose networks and evaluate the performance, the 50 annotated WSIs were split into 5 sets of 10 WSIs for cross validation. In each of the 5 folds, training was done on 30 WSIs (60%), validation on 10 WSIs (20%), and testing on 10 WSIs (20%). Annotated WSIs to construct the TDLU network were split into 9 sets of 10 or 11 WSIs for cross validation. For each of the 9 folds, training was done on 7 sets (~78%), validation on 1 set (~ 11%), and testing on 1 set (~ 11%). For all three methods, the model results from the first dataset split was used in all subsequent experiments (the models from the remaining folds are only used to evaluate the performance of the individual methods). All CNN models are described in the S1 Methods, and the acini detection network has been previously described [29]. To assess whether the training sets were large enough to learn to detect acini and segment TDLU ablation experiments were performed.
The three individual networks were integrated into a single automated method. This method can determine the three standardized quantitative measures by Figueroa et al. (i.e., TDLUs/mm², median TDLU span (μm), and median acini/TDLU [4, 10, 13, 14]) as well as two additional quantitative measures: number of acini per tissue area (acini/mm²) and median

![Fig 1. Examples of annotations for acini (A; annotated by blue squares), terminal duct lobular units (B) and adipose tissue (C).](https://doi.org/10.1371/journal.pone.0231653.g001)
TDLU area (mm²). Our method can also perform TDLU involution assessment using qualitative categories as described by Russo et al. [2] (i.e., predominant lobule type 1, 2 or 3) and Baer et al. [9] (i.e., no type 1 lobules, predominantly type 1 and no type 3, and mixed lobules (all others)). Thus, in total, our automated method can capture five quantitative and two qualitative measures of TDLU involution.

Validating the automated measures of TDLU involution

We validated our automated method by comparing automated results with manual assessment on an independent set of 40 WSIs (Table 1). Sixty WSIs were initially chosen at random from the NHS/NHSII BBD cases to contain 30 pre- and 30 post-menopausal women. Upon further review, we excluded one woman who had type 4 lobules which suggests that she was pregnant or lactating at time of BBD diagnosis. By excluding type 4 lobules, our method is generalizable to non-pregnant/not lactating women.

For manual assessment (n = 59 WSIs), two observers assessed the three standardized quantitative measures. Each observer randomly selected a ROI of approximately 50 mm² that contained an adequate number of normal TDLUs [4]. Within the ROI, the observers estimated the percentage of breast tissue (0 to 100%) and tissue containing adipose cells (<25%, 25–50%, 50–75%, or >75%), counted the total number of TDLUs, and randomly selected up to 10 normal TDLUs to measure span (μm) and count the number of spherical acini. TDLU boundary was defined by non-specialized/extra-lobular stroma. TDLUs were not counted if >50% of their acini were dilated by 2- to 3- fold, had metaplastic changes, or displayed ductal hyperplasia. TDLUs with <50% dilated acini were included and the acini within these TDLUs were counted (including dilated ones). Acini with elongated shape or no lumen were excluded. Three observers performed qualitative assessments using predominant lobule type by Russo et al. [2] and categories by Baer et al. [9]. For intra-observer evaluation, 10 out of 40 WSIs were randomly chosen for re-assessment.

Preliminary analyses of the 59 WSIs showed that although the manual and automated TDLU assessments were highly correlated, the values of the automated results for the number of acini per TDLU were lower than manual results. Therefore, we randomly selected 19 WSIs and linear regression to derive calibration weights based on the manual results to adjust our automated results. This calibration produced more meaningful values for interpretation. We applied the calibration weights to our automated results on the remaining 40 WSIs. The
calibration coefficient to adjust the automated number of acini per TDLU measure to the manual results was found to be 3.888. The intercept was not significantly different from zero. We applied the calibration coefficient to our automated results on the remaining 40 WSIs by multiplying all median number of acini per TDLU outcomes by 3.888.

Tissue area was adjusted for the percentage of adipose tissue by multiplying the total tissue area by the percentage of non-adipose tissue. Since manual observers only estimated adipose tissue percentage in categories (<25%, 25–50%, 50–75%, or >75%), we used the center bin values for this multiplication.

**Association of TDLU measures with age and menopausal status**

We also assessed manual and automated TDLU involution measures with age and menopausal status in the final 40 cases. This was to confirm that our measures were reflective of TDLU involution, as older women were expected to have more involution.

**Statistical analysis**

The evaluation of the acini detection neural network model was done using the F1 score and the evaluation of the TDLU and adipose tissue segmentation network models was done using the Dice similarity coefficient. F1 score is the harmonic mean of precision (i.e., sensitivity) and recall (i.e., positive predictive value), which assesses how accurate the automated detection compares with ground truth (i.e., manual annotation). The calculation for the Dice similarity coefficient is identical to F1 score, except it assesses the accuracy of the automated segmentation when compared to ground truth. The F1 score and Dice similarity coefficient are similar. Traditionally, when used to evaluate the detection performance this score is referred to as the F1 score and when used to evaluate the performance of a segmentation algorithm it is referred to as the Dice similarity coefficient.

Inter- and intra-observer agreements for quantitative measures were summarized using intraclass correlation coefficient (ICC). Two-way mixed effects, consistency, single rater (ICC (3,1)) was used. ICC values of <0.5, between 0.5 and 0.75, between 0.75 and 0.9, and >0.9 are indicative of poor, moderate, good, and excellent reliability, respectively [45]. Intra- and inter-observer agreements for qualitative measures were determined by Fleiss’ Kappa. For comparison with automated results, the consensus of the three observers was used. The consensus was determined by majority voting.

To determine the strength and direction of association of quantitative TDLU involution measures with age, Spearman’s rank correlation coefficient was used. The Kruskal-Wallis test was used to examine the differences between groups of qualitative measures and age. Mann-Whitney U and Chi-squared tests were used to assess the independence of quantitative and qualitative TDLU involution assessment with menopausal status. The scores for F1, Dice, and Fleiss’ Kappa range from 0 to 1, with 1 indicating perfect correlation. Analyses were performed using R and \( p < 0.05 \) was considered statistically significant. The ICC confidence intervals were calculated using the ICC function in the irr R package.

**Results**

**Performances of individual networks and establishing the automated method**

The F1 score of the acini detection method was 0.73±0.07 [29]. The TDLU and adipose tissue segmentation methods obtained Dice similarity coefficients of 0.84±0.13 and 0.87±0.04,
respectively. Ablation experiments showed that the methods converged with increasing number of training samples (S1 Fig).

Based on this quantitative evaluation, which indicates good agreement, and subsequent qualitative assessment we determined that the performances of these three networks were adequate to be integrated into one automated method (Figs 2 and 3; S2 Fig).

Fig 2. Results of the acini detection (A), terminal duct lobular unit (B), and adipose tissue (C) segmentation algorithms. The original images are in the left column, the middle column shows ground truth as annotated by human observers, and the detections and segmentations performed by the automated method are displayed in the right column.

https://doi.org/10.1371/journal.pone.0231653.g002
The primary cause of discordance between manual assessment and the automated method was the detection of acini and TDLU with proliferative or metaplastic changes which were intentionally excluded from manual annotation. For example, in S2C Fig, our method incorrectly segments intraductal papillomas as TDLUs despite correctly identifying other TDLUs.

Quantitative measures: Intra- and inter-observer agreement

Overall, quantitative measures derived from automated and manual methods achieved moderate to good inter-observer agreement (Table 2). The intra-observer agreement was good to excellent (ICC scores >0.8, 95% CI [0.53, 0.99]) and the inter-observer agreement among the two observers was moderate to good (ICC scores >0.7, 95% CI [0.51, 0.90]). The inter-observer agreement between the observers and the automated method was also moderate to good (ICC scores >0.5, 95% CI [0.19, 0.90]).

Qualitative measures: Intra- and inter-observer agreement

Qualitative measures between the three observers and the automated method achieved fair to moderate agreement (Table 3). Among the three observers, the inter-observer Kappa scores

| Table 2. Inter- and intra-observer intraclass correlation coefficient (ICC) scores and the 95% confidence interval (CI) for the quantitative terminal ductal lobular unit involution measures obtained from two observers and the automated method. |
|---------------------------------|------------------|-----------------|-----------------|-----------------|
|                                | Intra-observer ICC (95% CI) | Inter-observer ICC (95% CI) |
|                                | Observer 1          | Observer 2       | Observer 1 vs 2  | mean(observers) vs automated |
| Number of TDLUs per tissue area (mm²) | 0.96 (0.86, 0.99)  | 0.82 (0.78, 0.98) | 0.71 (0.51, 0.83) | 0.80 (0.63, 0.90) |
| Median TDLU span (μm)          | 0.91 (0.69, 0.98)  | 0.90 (0.67, 0.98) | 0.81 (0.67, 0.90) | 0.57 (0.19, 0.77) |
| Median number of acini per TDLU | 0.91 (0.69, 0.98)  | 0.86 (0.53, 0.96) | 0.73 (0.54, 0.85) | 0.80 (0.62, 0.89) |

*Intra-observer ICC was evaluated using 10 out of the 40 cases.
*Inter-observer ICC was evaluated using 40 cases.

https://doi.org/10.1371/journal.pone.0231653.t002

Fig 3. Results of the acini detection, terminal duct lobular unit, and adipose tissue segmentation algorithms (B) overlaid on the original image (A).

https://doi.org/10.1371/journal.pone.0231653.g003
were fair to moderate ($\kappa > 0.35 \ (p < 0.01)$) while there was a large variation in their intra-observer Kappa scores ($\kappa$ from 0.048 ($p = 0.880$) to 1.000 ($p < 0.01$)). The inter-observer agreement between the observers and the automated method was moderate ($\kappa > 0.5 \ (p < 0.01)$). There was slightly more agreement in the evaluation of Russo et al. [2] predominant lobule type compared to Baer et al. [9] categories.

### TDLU involution with age and menopausal status

All quantitative and qualitative measures obtained by manual and automated methods were significantly associated with age ($p < 0.05$; Figs 4 and 5). Table 4 shows the relationships between TDLU measures and menopausal status. All quantitative measures were significantly different between pre- and post-menopausal women, except number of TDLUs per tissue area evaluated by the automated method ($p = 0.06$). Likewise, qualitative measures (consensus vote by observers and automated method) were significantly different between pre- and post-menopausal women, except lobular classification according to Baer et al. [2] assessed by the automated method ($p = 0.07$). No participant was classified as predominantly type 3 according to Russo et al. [2]. Qualitative measures when assessed by individual observers were not associated with menopausal status ($p > 0.05$; S1 Table).

Thus, older and post-menopausal women had significantly fewer TDLUs/mm$^2$, smaller TDLUs, reduced number of acini per TDLU, and fewer acini/mm$^2$ compared to pre-menopausal women. Type 1 lobules were predominantly observed in post-menopausal women while the majority of pre-menopausal women had mixed lobules.

### Discussion

Greater amounts of TDLU involution are inversely associated with breast cancer risk [5, 6, 9–11] and aggressive breast cancer subtypes [13, 14]. It is important to better understand TDLU involution as well as epidemiological factors that influence the involution process to obtain deeper insights into breast carcinogenesis and identify new opportunities for breast cancer prevention. A major bottleneck to studying TDLU involution and breast cancer risk in large epidemiological cohorts is the need for manual qualitative and/or quantitative assessment by pathologists. In this study, we developed and validated a computational pathology method that can assess five quantitative and two qualitative measures of TDLU involution. Our automated method was highly comparable to manual assessment, and we confirmed that our TDLU involution measures reflect age and menopausal status [4]. This technology will be a valuable research tool to facilitate future breast cancer risk studies.
Fig 4. Scatterplots of the association of quantitative terminal ductal lobular unit (TDLU) involution measures and age. TDLU count per tissue area assessed using manual (A) and automated (B) method were significantly inversely correlated with age ($p<0.01$). Median TDLU span assessed manually (C) and with the automated method (D) was significantly inversely correlated with age ($p<0.01$ and $p=0.01$). Median acini count per TDLU assessed using manual (E) and automated (F) assessment was also significantly inversely correlated with age ($p<0.01$). Acini count per tissue area assessed by the automated method was significantly inversely correlated with age (G; $p<0.01$). Median TDLU area assessed by the automated method was significantly inversely correlated with age (H; $p<0.01$).

https://doi.org/10.1371/journal.pone.0231653.g004
Our automated method integrates three separate networks for acini detection, TDLU segmentation, and adipose tissue segmentation. It was challenging to develop the TDLU segmentation network compared to the other two networks because TDLUs have highly variable appearances and BBD encompasses a wide range of morphology. As such, the TDLU segmentation network required more training WSIs to achieve a Dice score similar to the adipose tissue segmentation network. Since we are the first to develop networks for acini detection and TDLU segmentation, we were unable to benchmark our networks. We identified three primary causes of discordance between manual assessment and the automated method which affected our F1 and Dice scores: 1) acini with proliferative or metaplastic changes were frequently detected by the network but were intentionally excluded from manual annotation; 2) the network had difficulty predicting boundaries of TDLUs with complex clustering; and 3) in some cases, the network interpreted large ducts as adipose tissue.

Despite researchers’ best efforts to create a perfect method, most automated methods remain prone to segmentation errors. Solutions to address these issues and improve our computational method include increasing the number of training samples with improved annotation and applying hard negative mining. The inclusion of abnormal epithelium when assessing TDLU involution may influence breast cancer risk assessment. Therefore, future work will evaluate the inter-variability of TDLU measures between slides obtained from different tissue blocks for each patient. In addition, summarizing the automated results using median instead of mean, and evaluating at least two WSIs per case (averaging the median values), will improve the robustness and reliability of the data in future studies. This study focuses...
on assessing TDLU involution in non-malignant breast tissue only. If this method were to be used to assess TDLU involution in tumor-adjacent normal breast tissues, it would need to be re-trained to include malignant tissue.

To capture TDLU span, the automated method uses the length of the major axis of the ellipse that is identical to the normalized second central moments for each TDLU. In contrast, a pathologist has to select two opposite points along the boundary of a TDLU to obtain the longest span. Thus, the manual assessment of TDLU span inevitably contains some subjectivity and explains the low inter-observer agreement score between manual and automated results.

Our automated method has the ability to capture two new measures: number of acini per tissue area and median TDLU area. Future studies will evaluate and compare these newer measures with the existing three standardized measures to determine which TDLU involution quantitative measure is most associated with breast cancer risk.

TDLU involution is historically assessed using qualitative measures [2, 5, 9]. The large variation in intra- and inter-observer Kappa scores as observed in this study reiterated the high

<table>
<thead>
<tr>
<th>Table 4. The association of terminal ductal lobular unit (TDLU) involution measures and menopausal status.</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Quantitative measures</td>
</tr>
<tr>
<td>Number of TDLU per tissue area (mm$^2$), median $n$ (IQR)</td>
</tr>
<tr>
<td>Evaluated by observers</td>
</tr>
<tr>
<td>Evaluated by the automated method</td>
</tr>
<tr>
<td>Median TDLU span in μm, median $n$ (IQR)</td>
</tr>
<tr>
<td>Evaluated by observers</td>
</tr>
<tr>
<td>Evaluated by the automated method</td>
</tr>
<tr>
<td>Number of acini per TDLU, median $n$ (IQR)</td>
</tr>
<tr>
<td>Evaluated by observers</td>
</tr>
<tr>
<td>Evaluated by the automated method</td>
</tr>
<tr>
<td>Number of acini per tissue area (mm$^2$), median $n$ (IQR)</td>
</tr>
<tr>
<td>Evaluated by the automated method</td>
</tr>
<tr>
<td>Median TDLU area (mm$^2$), median $n$ (IQR)</td>
</tr>
<tr>
<td>Evaluated by the automated method</td>
</tr>
<tr>
<td>Qualitative assessment</td>
</tr>
<tr>
<td>Predominant lobular type by observers (consensus vote), n (%)</td>
</tr>
<tr>
<td>Type 1</td>
</tr>
<tr>
<td>Type 2</td>
</tr>
<tr>
<td>Type 3</td>
</tr>
<tr>
<td>Predominant lobular type by the automated method, n (%)</td>
</tr>
<tr>
<td>Type 1</td>
</tr>
<tr>
<td>Type 2</td>
</tr>
<tr>
<td>Type 3</td>
</tr>
<tr>
<td>Lobular classification according to Baer et al. [2] by observers (consensus vote), n (%)</td>
</tr>
<tr>
<td>No type 1</td>
</tr>
<tr>
<td>Mixed lobules</td>
</tr>
<tr>
<td>Predominantly type 1, no type 3</td>
</tr>
<tr>
<td>Lobular classification according to Baer et al. [2] by the automated method, n (%)</td>
</tr>
<tr>
<td>No type 1</td>
</tr>
<tr>
<td>Mixed lobules</td>
</tr>
<tr>
<td>Predominantly type 1, no type 3</td>
</tr>
</tbody>
</table>

https://doi.org/10.1371/journal.pone.0231653.t004
subjectivity of qualitative measures, thus spurring researchers to develop standardized quantitative measures to assess TDLU involution [4, 10, 13, 14]. Our study showed higher intra- and inter-observer agreement for quantitative measures compared to qualitative measures, again highlighting the reproducibility of quantitative measures. Despite assessing different tissue areas for manual assessment (observers selected 50 mm² tissue area) and automated method (entire tissue area on WSI), the good agreement between the observers and automated results provided additional assurance that our automated method is comparable to manual assessment.

Conclusion
We developed and validated an automated method to measure TDLU involution as a first step towards automated prediction of breast cancer risk. Qualitative assessment of TDLU involution is a subjective process. Quantitative assessment produces more reproducible results but is labor-intensive for pathologists. Our method can eliminate the labor-intensiveness and subjectivity of manual TDLU involution assessment. Our technology can be applied on a larger scale to assess breast cancer risk in epidemiological studies. Future work will determine the best quantitative TDLU involution measure to predict breast cancer risk, and evaluate the impact of incorporating these measures into clinical breast cancer risk assessment models to improve patient management.

Supporting information
S1 Fig. Line charts demonstrating the F1 score obtained on the test set with models trained using different percentages of the training dataset. (A) The ablation experiment for the detection of acini. The line converges before it reaches 100% of the training data indicating that the training set is large enough. (B) The ablation experiment for the segmentation of TDLUs. The line converges before it reaches 100% of the training data indicating that the training set is large enough. The line charts show the mean value and standard deviation. (DOCX)

S2 Fig. Results of the automated method (A.2, B.2, C.2) overlaid on original whole slide images (A.1, B.1, C.1). Detected acini are shown in blue, terminal duct lobular units (TDLUs) in pink, and adipose tissue in yellow. The black crosses (C.2) indicate regions where intraductal papillomas were incorrectly segmented as TDLUs. (DOCX)

S1 Table. The association of terminal ductal lobular unit (TDLU) involution measures and menopausal status. (DOCX)

S1 Methods. (DOCX)

Acknowledgments
We would like to thank the participants and staff of the Nurses’ Health Study and Nurses’ Health Study II for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY. The authors assume full responsibility for analyses and interpretation of these data.
Author Contributions


Formal analysis: Suzanne C. Wetstein, Yujing J. Heng.

Funding acquisition: Josien P. W. Pluim, Rulla M. Tamimi, Yujing J. Heng, Mitko Veta.

Investigation: Suzanne C. Wetstein.

Methodology: Suzanne C. Wetstein, Bart Bakker, Ruud Vlutters, Marinus B. van Leeuwen, Josien P. W. Pluim, Yujing J. Heng, Mitko Veta.

Project administration: Rulla M. Tamimi, Yujing J. Heng.

Resources: Bart Bakker, Ruud Vlutters, Marinus B. van Leeuwen.

Supervision: Josien P. W. Pluim, Yujing J. Heng, Mitko Veta.

Validation: Suzanne C. Wetstein.

Visualization: Suzanne C. Wetstein.

Writing – original draft: Suzanne C. Wetstein, Allison M. Onken, Christina Luffman, Gabrielle M. Baker, Michael E. Pyle, Kevin H. Kensler, Ying Liu, Bart Bakker, Ruud Vlutters, Marinus B. van Leeuwen, Laura C. Collins, Stuart J. Schnitt, Josien P. W. Pluim, Rulla M. Tamimi, Yujing J. Heng, Mitko Veta.


References


