Data for Breast Cancer Vaccine Research

Evidence reviewed through 8 September 2026

Data requirements and research resources for antigen discovery functional validation and clinical development

This report is a narrative review of evidence and data resources drawn from publicly available primary sources. It was prepared with AI-assisted searching and organization, has not been peer reviewed, and includes no new human experiments or original patient-data analysis. It complements the earlier breast cancer vaccine frontier review. It is not a substitute for a clinical protocol, ethics review, regulatory consultation, or individual medical advice.

Core conclusions

The most important gap in breast cancer vaccine development is data that connects several questions within the same patient: which antigens the tumor contains, which are actually presented, whether immune cells recognize and kill the tumor, what happens after treatment, and whether the cancer recurs years later. The useful unit of research links a patient, timepoint, lesion and specimen to antigen evidence, immune responses, treatment, and outcomes. An isolated image or report cannot provide that chain of evidence.

Existing breast cancer vaccine studies support this paired approach, but do not establish a cure. A 2026 Nature study of individualized mRNA vaccination observed durable T-cell responses and illustrated how recurrence can accompany impaired antigen presentation. It was a small, nonrandomized study and cannot by itself establish that vaccination reduces recurrence or prolongs survival.[1] A 2024 Genome Medicine study of individualized DNA vaccination in triple-negative breast cancer integrated tumor and normal sequencing, candidate antigen design, and immune validation, and reported sequencing resources available through an access application.[2]

The priorities in this report are an interpretation for research planning, not a consensus-based quantitative ranking. The first is evidence of actually presented tumor antigens and their safety in normal tissues. The second is antigen-matched functional immunity and explanations for failure. The third is prospective evidence, preferably randomized, linking these measurements to long-term outcomes alongside modern standard treatment. Imaging, pathology, electronic health records, and patient reports matter, but cannot replace these evidence categories.

Search scope and evidence quality

The targeted search covered original research in the Nature journals, Cell, JCI, Genome Medicine, Cancer Discovery, and PNAS; ClinicalTrials.gov registrations and results; final and draft FDA guidance; and providers including GDC, dbGaP, GEO, PRIDE, CEDAR, TCIA, and HTAN. It emphasized studies from 2024 to 2026 while retaining older foundational datasets. Checks included data-availability statements, registry updates, accession identifiers, access routes, study populations, and licensing limits.

1 Sahin U, Schmidt M, Derhovanessian E, et al (2026). Full citation and original link in Sources, entry 1.

2 Zhang X, Goedegebuure SP, Chen MY, et al (2024). Full citation and original link in Sources, entry 2.

This is a broad, targeted research map, not an exhaustive PRISMA systematic review. It does not claim to cover every paper or unpublished company dataset. An available resource means its entry point and access route were checked, not that access to controlled raw files was approved. A resource listing is not permission for commercial use. Human clinical evidence, human association studies, methods developed in other cancers, animal studies, and regulatory documents are distinguished throughout.

1 Define the vaccine development setting

Preventing a first breast cancer and preventing recurrence after treatment are different development tasks. FDA guidance on therapeutic cancer vaccines addresses therapeutic development in people with a cancer history, not primary prevention in people without one. It is not a universal experimental procedure for every vaccine platform.[3]

Development setting	Central data question	Essential clinical evidence
Primary prevention	Baseline risk in high-risk people without cancer; antigen differences between precancerous and normal tissue; long-term and reproductive safety	Incidence and benefit-risk over adequate follow-up, rather than increases in antibodies or T cells alone
Prevention of recurrence	Antigens in primary and residual lesions; minimal residual disease; immune memory; standard treatment exposure	Recurrence, distant recurrence, and survival outcomes with appropriate controls
Metastatic disease treatment	Antigens and immune escape in current metastases; prior treatment; variation across organs and lesions	Durable disease control, clinical benefit, and survival rather than a single immune signal

The platform also determines data requirements. Personalized neoantigen vaccines rely heavily on paired patient sequencing and candidate prioritization. Shared-antigen vaccines require population coverage and normal-tissue safety evidence. Cell-based vaccines also require cell identity, viability, and functional quality control. Not every patient needs every expensive assay listed below. Discovery cohorts, validation cohorts, and clinical trials must collectively provide adequate evidence across the relevant layers.

2 Data requirements across development

The following is a research data-dictionary framework, not a list of tests patients should purchase. A requirement means sufficient evidence at the relevant development stage, not mandatory collection of every measurement from every participant.

An antigen is a molecular feature the immune system can recognize. Human leukocyte antigen molecules, or HLA, display peptide fragments to immune cells. T-cell receptors, or TCRs, recognize these signals. WES and WGS mean whole-exome and whole-genome sequencing. Circulating tumor DNA, or ctDNA, is tumor-derived DNA in blood; MRD refers to minimal or molecular residual disease. DFS, EFS, and OS mean disease-free, event-free, and overall survival. Event definitions must follow the trial protocol.

3 US Food and Drug Administration (2011). Full citation and original link in Sources, entry 3.

Data layer	Core information to retain	Question addressed and limitations
1 Clinical and pathological stratification	Age, diagnosis date, histology, stage, ER/PR/HER2, grade, residual cancer burden, lesion site, and genetic risk where relevant	Keep TNBC, HER2-positive, and hormone receptor-positive disease distinct; distinguish primary disease from recurrence
2 Paired tumor and normal DNA	WES or WGS; single-nucleotide variants, indels, fusions, copy number, clonality, purity, depth, and quality control	Separate somatic from germline variants and identify antigen sources; a genetic test summary is not a complete analytical input
3 Tumor RNA	Expression, mutant allele expression, transcripts, fusions, and splicing; Ribo-seq in some studies	DNA presence does not establish expression; RNA expression does not establish translation or immune visibility
4 HLA and presentation pathways	High-resolution HLA class I and II typing; expression and loss of heterozygosity; B2M, TAP, and related pathways	Assess possible peptide presentation and whether the tumor has disabled presentation
5 HLA immunopeptidomics	HLA-bound peptide mass spectrometry, raw spectra, search libraries, false-discovery rates, sequence origins, and allele assignment	Move from predicted to observed presentation; lack of detection does not prove absence
6 Normal tissue safety	RNA, protein, and HLA peptides across normal or benign organs; relevant physiological and inflammatory states	Assess cross-recognition and potential off-target risk; high tumor expression alone is insufficient
7 Functional immunity	Antigen-specific CD4/CD8 responses, ELISpot/ICS, multimers, paired TCRs, recognition and killing of actual tumor cells, and antibody function where relevant	HLA binding, an immune response, and tumor killing are different evidence levels
8 Tumor microenvironment	Single-cell RNA/TCR, spatial RNA/protein, immune neighborhoods, exhaustion and memory states, and suppressive cells	Explain why strong blood responses may fail to enter the tumor or function within it
9 Longitudinal lesions and specimens	Tissue and blood before and during treatment, after surgery, and at recurrence or metastasis; timepoint and lesion identifiers	Identify antigen loss, subclonal selection, and immune escape; an old primary tumor may not represent a current metastasis
10 ctDNA and MRD	Serial quantitative results, detection limits, assay and version, sampling times, trackable variants, and false-positive assessment	Distinguish residual disease, clearance, and reappearance; a negative test does not establish tumor absence
11 Imaging and digital pathology	Original DICOM, associated reports, lesion annotations, acquisition parameters, whole-slide pathology, and raw or quantitative IHC/ISH	Establish lesions, burden, and outcomes; radiology images and pathology slides are different data formats
12 Treatment and product	Standard and concomitant treatment dates, doses, interruptions; vaccine sequence or construct version, batch, potency, stability, manufacturing failures, and delivery times	Separate vaccine effects from standard treatment and explain product, logistical, or immune failures

Data layer	Core information to retain	Question addressed and limitations
13 Clinical endpoints and safety	Randomization, event definitions, DFS/iDFS/EFS, distant recurrence, OS, adverse events, quality of life, last follow-up, and loss to follow-up	Demonstrate clinical benefit; immune signals or one imaging assessment cannot replace long-term controlled outcomes
14 Provenance and governance	Consent scope, research and commercial permissions, patient-specimen links, original file sources, experimental batches, missingness reasons, and analysis versions	Support traceability, reproducibility, and authorized use; unmeasured is not negative

This framework draws on the vaccine studies, immunopeptidomic resources, immune-function benchmarks, longitudinal cohorts, and regulatory materials discussed below. Section 4 lists access routes for the relevant data layers.

Three gaps in target discovery

First, evidence of actual presentation is often missing. A JCI breast cancer study identified HLA-presented peptides in 26 primary tumors and selected tumor-specific or tumor-associated antigens. Only one of its 25 tumor-specific antigens came from a mutation; the others arose from aberrant expression. Candidate discovery should therefore extend beyond point mutations, but these candidates are not clinically validated vaccine targets.[4]

Second, normal-tissue references are incomplete. HLA Ligand Atlas provides a normal and benign tissue peptide-presentation reference that can help exclude candidates lacking tumor specificity. Its donors, tissues, and physiological states are limited. Failure to find a candidate in the database cannot establish its safety.[5]

Third, high-quality experimental failure labels are scarce. The TESLA consortium experimentally compared predictions from several teams using data from six patients with melanoma or lung cancer. Of 608 tested candidates, 37 were classified as immunogenic. This is a cross-cancer methodological benchmark, not a breast cancer response rate. Its lesson is to retain failed candidates together with the HLA, assay, and detection threshold under which failure was observed, rather than saving only positive examples.[6]

Biomarkers cannot replace clinical outcomes

ctDNA is a useful research tool, but ctDNA clearance cannot simply be called a cure. The FDA's 2024 final guidance emphasizes prospective randomized trials and long-term outcomes for validating particular ctDNA uses, with attention to methods, sampling times, and analytical performance.[7] In the breast cancer c-TRAK TN study, some patients allocated to intervention already had metastases on staging after ctDNA detection. Research must therefore establish whether a genuinely actionable window exists, not only whether earlier detection is possible. This was not a vaccine trial.[8]

4 Kina E, Laverdure JP, Durette C, et al (2024). Full citation and original link in Sources, entry 4.

5 Marcu A, Bichmann L, Kuchenbecker L, et al (2021). Full citation and original link in Sources, entry 5.

6 Wells DK, van Buuren MM, Dang KK, et al (2020). Full citation and original link in Sources, entry 6.

7 US Food and Drug Administration (2024). Full citation and original link in Sources, entry 7.

8 Turner NC, Swift C, Jenkins B, et al (2023). Full citation and original link in Sources, entry 8.

Product data are also easily overlooked. An identical antigen list does not guarantee identical biological activity across batches. In August 2026, the FDA issued draft guidance on potency assessment of active immunotherapy products, highlighting the value of connecting potency assays to mechanism of action early in development. It remained draft guidance at the evidence cutoff and should not be described as an implemented mandatory standard.[9]

HLA coverage requires direct typing and population data. IPD-IMGT/HLA provides standardized nomenclature and reference sequences.[10] Allele Frequency Net Database supports assessment of HLA frequencies across populations. Frequency coverage is not efficacy coverage, and race or nationality cannot substitute for an individual's HLA typing.[11]

3 Additional evidence for primary prevention

Preventing breast cancer in people who have never had it requires evidence in high-risk populations and relevant normal physiological states. Results in patients after cancer treatment cannot simply be transferred to this setting.

A September 2026 PNAS study of Lalba mRNA-LNP vaccination suggests a preventive antigen direction, but used a rat mammary tumor model. It does not establish protection in healthy women.[12] A Nature study of breast tissue-resident T cells after pregnancy and lactation combines animal mechanisms with human observations. It supports studying reproductive history, local immune memory, and tissue state; it is not a recommendation to have children to prevent or treat cancer.[13]

A 2026 Cancer Discovery spatial study examined the transition from precancer to cancer across contiguous tissue regions. Its 51 cases included both breast and prostate cancer; they were not 51 breast cancer cases. Nor was it a natural-history study following the same healthy women for years.[14] The alpha-lactalbumin vaccine registration NCT04674306 describes exploration in high-risk groups and conditional data sharing, but does not establish protective efficacy for the general population of healthy women.[15]

Additional data need	Why primary prevention requires it	Insufficient substitute
Absolute incidence risk and long-term natural history	Identify who may benefit and how many events follow-up must capture	Family history, self-reported risk, or a single screening examination alone
Paired antigens in precancerous benign and normal breast tissue	Identify targets present early in malignant transformation and as absent as possible in normal tissue	Antigen libraries derived only from advanced tumors

9 US Food and Drug Administration (2026). Full citation and original link in Sources, entry 9.

10 IPD-IMGT/HLA Database (2026). Full citation and original link in Sources, entry 10.

11 Allele Frequency Net Database (2026). Full citation and original link in Sources, entry 11.

12 Nishida J, Seehawer M, Rojas Jimenez E, et al (2026). Full citation and original link in Sources, entry 12.

13 Virassamy B, Caramia F, Savas P, et al (2026). Full citation and original link in Sources, entry 13.

14 Storrs E, Mo CK, Chou WH, et al (2026). Full citation and original link in Sources, entry 14.

15 ClinicalTrials.gov, NCT04674306 (2026). Full citation and original link in Sources, entry 15.

Additional data need	Why primary prevention requires it	Insufficient substitute
Reproductive and physiological stage	Relevant proteins and local immunity may vary with pregnancy, lactation, and menopause	Treating absence in one normal sample as lifelong absence
Long-term safety and immune persistence	People without cancer have lower tolerance for toxicity, immune abnormalities, and delayed harm	Short phase 1 tolerability or a few blood immune measurements
Risk-stratified controls and incidence endpoints	Measure prevention benefit and avoid overtreatment	Delayed tumor onset in animals or absence of recurrence after treatment

Additional studies should address when precancerous antigens appear, normal-tissue cross-reactivity, HLA and risk-group coverage, local and systemic immune memory, long-term safety, and actual differences in cancer incidence. Additional tissue collection must follow ethical review and clinical justification; a platform's desire for complete data does not justify unnecessary biopsies.

4 Available data resources

The resources below proceed from direct vaccine and antigen data to molecular, spatial, imaging, and clinical resources. Open, controlled-access, and collaboration-dependent data require separate planning. Public databases can often support initial discovery or methods development. However, this search did not identify an open patient-level breast cancer vaccine dataset that completely connects antigens, functional experiments, manufacturing, randomization, and long-term outcomes.

A Direct vaccine and antigen data

The FR-alpha vaccine study NCT03012100 illustrates an important gap. Its registry provides aggregate randomized clinical results, but as of the 30 June 2026 update, relevant antigen-level and vaccine-induced T-cell outcome measures were still marked NOT_POSTED. That does not prove the data were never collected. It does mean external researchers cannot reconstruct mechanisms of success or failure from published clinical summaries alone. The DFS comparison did not show a statistically significant benefit and should not be promoted as successful recurrence prevention.[16]

Resource and entry point	Available information	Access and important limits
BioNTech individualized mRNA TNBC study Nature article [1]	Paper, supplements, immune analyses, and follow-up; an example of paired study design	No complete open package of linked patient-level raw sequencing and clinical data was verified; follow the authors' materials and collaboration terms
WashU DNA vaccine dbGaP phs002787.v2.p1	18 consenting participants; metadata explicitly lists paired tumor/normal exomes and RNA sequencing	Controlled application; v2 supersedes v1. Confirm TCR and other reported modalities against the approved file inventory rather than assuming completeness [2]

16 ClinicalTrials.gov, NCT03012100 (2026). Full citation and original link in Sources, entry 16.

Resource and entry point	Available information	Access and important limits
Breast cancer immunopeptidomics PXD034818 [4]	Raw mass spectrometry and related results; RNA entry PRJNA852282; transcript data GSE206838	A direct starting point for actual presentation in breast cancer, not vaccine clinical efficacy; check companion identifiers, patents, and usage rights
TESLA Cell paper and supplements [6]	Neoantigen prediction inputs, experimental validation, and negative-candidate benchmarks	Melanoma and lung cancer, not breast cancer; do not directly transfer performance estimates
FR-alpha vaccine NCT03012100 [16]	Aggregate randomized clinical and safety results and prespecified endpoint information	Aggregate results are not raw participant data; immune correlates are not fully public
Alpha-lactalbumin vaccine NCT04674306 [15]	Registration, design, and sharing plan to identify possible collaborations	Individual data sharing is conditional for named institutions or potential pharmaceutical partners, with confidentiality agreements; not unrestricted download

The WashU entry was checked against the current dbGaP v2 registration, which explicitly lists exome, RNA, and paired tumor-normal study types.[17]

CEDAR is a more focused starting point for cancer antigens than an undirected IEDB search. It integrates cancer epitope, antibody, T-cell, HLA-ligand, and receptor evidence. Its August 2026 provider-authored paper describes the resource; it is not a new breast cancer efficacy trial.[18] At the evidence cutoff, the export page listed an update dated 1 September 2026 and a CC BY 4.0 license. Users should still retain the download version, original study, and experimental conditions.[19]

Resource and entry point	Available information	Access and important limits
CEDAR database downloads	Cancer epitopes, T/B-cell experiments, HLA ligands, receptors, and CSV/TSV exports	Supports candidates and experimental labels, not years of patient history; measurements from different papers are not automatically comparable
HLA Ligand Atlas [5]	HLA class I and II presented peptides from normal or benign tissues	Portal indicates CC BY; coverage is limited and non-detection is not proof of safety
IPD-IMGT/HLA [10]	Versioned HLA nomenclature and reference sequences	Not a patient HLA cohort or population efficacy dataset

17 NCBI dbGaP (2026). Full citation and original link in Sources, entry 17.

18 Koşaloğlu-Yalçın Z, Carri I, Marrama D, et al (2026). Full citation and original link in Sources, entry 18.

19 Cancer Epitope Database and Analysis Resource (2026). Full citation and original link in Sources, entry 19.

Resource and entry point	Available information	Access and important limits
Allele Frequency Net [11]	Population HLA allele and haplotype frequencies and quality labels	Populations, sample sizes, and quality vary; coverage assessment does not replace typing or clinical validation

B Tumor molecular and normal tissue resources

GDC separates open derived data from controlled data. A visible TCGA project page does not mean every patient's raw sequences are available for immediate download.[20] CPTAC breast cancer proteogenomics connects DNA, RNA, and protein measurements, but a conventional proteome is not an HLA-presented peptidome.[21] METABRIC supports expression, copy-number, and historical outcome analyses. Coverage varies by modality, so the portal's total enrollment is not a count of complete multi-omic cases.[22]

Resource and entry point	Available information	Access and important limits
TCGA-BRCA/GDC	Clinical, biospecimen, expression, mutation, copy-number, some slide, and multi-omic information	Some derived data are open; raw or sensitive genomic data are controlled. Vaccine exposure and standardized antigen-functional outcomes are lacking
CPTAC/PDC PDC000120	Breast cancer protein, phosphorylation, and related omics with corresponding genomic resources	Proteomic and raw genomic permissions differ; not direct HLA immunopeptidomics
METABRIC/ cBioPortal	Expression, copy number, mutations, and long-term clinical information	Uneven modality coverage and historical treatment; not an equivalent control for a modern vaccine trial

AURORA US links primary and metastatic tumor multi-omics and is particularly relevant to whether targets in a primary tumor persist in metastases. Its 55 women contributed multiple lesions and specimens, and not every case has every assay. Controlled and open data are distributed between dbGaP and GEO.[23] The Wu breast cancer single-cell and spatial atlas supports immune microenvironment analyses.[24] The Kumar adult breast atlas adds cell-type and normal-tissue context, but specimen sources and donor backgrounds must be retained. Not every normal specimen should be treated as a healthy control without any disease-related context.[25]

20 National Cancer Institute Genomic Data Commons (2026). Full citation and original link in Sources, entry 20.

21 Krug K, Jaehnig EJ, Satpathy S, et al (2020). Full citation and original link in Sources, entry 21.

22 cBioPortal (2026). Full citation and original link in Sources, entry 22.

23 Garcia-Recio S, Hinoue T, Wheeler GL, et al (2023). Full citation and original link in Sources, entry 23.

24 Wu SZ, Al-Eryani G, Roden DL, et al (2021). Full citation and original link in Sources, entry 24.

25 Kumar T, Nee K, Wei R, et al (2023). Full citation and original link in Sources, entry 25.

Resource and entry point	Available information	Access and important limits
AURORA US study	Primary/metastatic multi-omics; dbGaP phs002622.v1.p1; RNA GSE209998; methylation GSE212375	Patient, lesion, and complete-pair counts differ; do not substitute European AURORA accessions
Wu breast cancer atlas GSE176078	Processed single-cell data; SCP1039; spatial data at Zenodo 10.5281/zenodo.4739739	Open processed and controlled raw data are separate; no complete vaccine-intervention outcomes
Adult breast atlas GSE195665	Subseries GSE235326 single-cell, GSE234817 single-nucleus, and GSE234814 spatial data	Normal breast cell-composition reference, not a substitute for normal HLA presentation measurements
GTEx adult downloads	Normal tissue gene/transcript expression and related resources	Open and controlled tiers; variation in batches, processing, and donors; not a patient's own matched normal sample
Human Protein Atlas	Normal protein, IHC, RNA, and cell-type references	License page specifies CC BY 4.0 and third-party rights; expression does not establish presentation or safety

GTEx and Human Protein Atlas broaden normal-tissue context and candidate filtering. They cannot prove that an antigen occurs exclusively in cancer cells.[26,27]

C Spatial immunity and experimental models

HTAN provides multicenter specimen, clinical, spatial, and multimodal resources with linkage standards. Open processed data and controlled raw genomic data are managed separately. The resource is also useful for learning how to connect patients, specimens, images, and molecular measurements.[28]

The 2026 TONIC spatial study included 103 women and 270 tumor samples and examined immune checkpoint inhibitor responses in metastatic TNBC, not vaccine efficacy. Processed images are available, but the sequencing and response-data statement restricts use to academic purposes; commercial permission cannot be assumed.[29] The contemporary QUICHE study provides spatial immune-neighborhood methods and imaging data. These are potential tools for vaccine mechanism research, not grounds for relabeling immunotherapy response as vaccine response.[30]

The August 2026 Nature HCMI compendium expands patient-derived experimental models. Its total spans multiple cancers and should not be described as a breast-only collection. Models lose or alter elements of the original immune and stromal environment, so they cannot alone establish human immunity or normal-tissue toxicity.[31]

26 GTEx Consortium (2026). Full citation and original link in Sources, entry 26.

27 Human Protein Atlas (2026). Full citation and original link in Sources, entry 27.

28 Human Tumor Atlas Network (2025). Full citation and original link in Sources, entry 28.

29 Greenwald NF, Nederlof I, et al (2026). Full citation and original link in Sources, entry 29.

30 Ranek JS, Greenwald NF, Goldston M, et al (2026). Full citation and original link in Sources, entry 30.

31 El Harouni D, Al-Jazrawe M, Choi S, et al (2026). Full citation and original link in Sources, entry 31.

Resource and entry point	Available information	Access and important limits
HTAN data portal	Single-cell, spatial, imaging, clinical, and specimen metadata; controlled parent study phs002371	Not every patient has every modality; link using patient-specimen identifiers, not disease labels alone
TONIC images S-BIAD1288	Multiplex protein images and processed data; analysis results at Zenodo 10.5281/zenodo.14112852	Public images do not establish commercial rights to all sequence/response data; apply the authors' conditions
QUICHE images S-BIAD1507	Spatial images; anndata 10.5281/zenodo.14290163; related NeoTRIP data 10.5281/zenodo.7990870	Each package has its own license; validate on held-out patients and centers, avoiding image-tile leakage
HCMI searchable catalog	Patient-derived models and linked clinical/genomic entry points	Material access involves terms, costs, and transfer agreements; models lack the complete native immune system

D Imaging and clinical trajectories

I-SPY2 imaging provides serial MRI and some treatment and pathological outcomes for treatment timelines and imaging-response studies. The collection page specifies CC BY 4.0.[32] Duke breast MRI provides preoperative imaging and associated clinical and pathological information, but specifies CC BY-NC 4.0. Commercial projects cannot treat it as unrestricted data.[33]

Resource and entry point	Available information	Access and important limits
TCIA I-SPY2	Multitimepoint MRI DICOM, segmentations, and clinical tables; the 985 people in Cohort 1 include 266 ACRIN 6698 participants	Avoid double counting; an imaging collection does not open every trial assay or biospecimen; not a vaccine cohort
TCIA Duke Breast Cancer MRI	922 preoperative MRI cases with clinical, pathological, treatment, and outcome information	Retrospective, single-center, historical, and noncommercially licensed; not a complete longitudinal EHR
ImmPort shared resources	Immune studies, clinical associations, flow cytometry, and data-organization standards	Check each study's cancer type, fields, and terms; not a ready-made complete breast cancer vaccine database

ImmPort can inform immune-assay data organization and identify relevant studies. Study design must be checked individually; an immunology classification alone does not make a dataset relevant to breast cancer vaccines.[34]

32 The Cancer Imaging Archive (2026). Full citation and original link in Sources, entry 32.

33 The Cancer Imaging Archive (2026). Full citation and original link in Sources, entry 33.

34 ImmPort (2026). Full citation and original link in Sources, entry 34.

5 Research priorities building on existing studies

Study 1 Identify safe antigen combinations with useful coverage

Core inputs are paired breast tumor/normal DNA, RNA, HLA, presented peptides, and normal-tissue references. The question is which antigens persist in enough malignant cells, are presented and potentially recognized across the intended HLA population, and minimize normal-tissue harm. A high peptide prediction score alone is insufficient. Shared-antigen approaches need coverage across subtypes and populations; personalized approaches need coverage across a patient's tumor clones and multiple targets.

A 2026 Nature study expanding the human proteome broadens the search space to noncanonical open reading frames, microproteins, and peptide sources. These sources also occur in normal tissues. Noncanonical does not mean tumor-specific. They should extend the candidate set for validation rather than serve as a ready-made safety-qualified target library for human products.[35]

The proposed output is an antigen table with evidence grades: sequence origin, mutation or expression support, patient and lesion, HLA, mass-spectrometry support, normal-tissue matches, functional assays, coverage, unknowns, and exclusion reasons. Screening thresholds and development decisions require joint input from qualified experimental and clinical teams.

Study 2 Predict candidates that generate functional antitumor responses

Tools such as pVACtools integrate variants, expression, HLA, and candidate prioritization. They are components of an analytical workflow, not proof of target validity.[36] Expression, presentation, immune recognition, and recognition of actual tumor cells should be predicted and evaluated as separate labels, not collapsed into a generic success label.

Training data should distinguish experimental negatives, untested candidates, assay failures, and missing values. Evaluation should split by patient and then use external validation across centers, time periods, or HLA groups. Splitting different peptides from the same patient, or tiles from the same tissue, across training and test sets does not demonstrate generalization. TESLA offers methodological guidance, but new models still need independent breast cancer validation.[6]

Study 3 Explain recurrence despite an immune response

Connect peripheral immunity, the spatial tumor microenvironment, changing antigen expression and presentation, HLA loss, tumor clones, and treatment exposure. Before-and-after specimens should come from comparable lesions, or their site changes should be explicit. Distinguish recurrence of the same tumor from a new primary cancer or a new lesion at another site.

The BioNTech, AURORA US, and TONIC studies offer different evidence on vaccine responses, metastatic evolution, and spatial context.[1,23,29] Together they can support hypotheses, but different patients in different studies cannot be assembled into an invented complete clinical history. New paired validation cohorts would provide the missing evidence.

35 Deutsch EW, Kok LW, Mudge JM, et al (2026). Full citation and original link in Sources, entry 35.

36 Hundal J, Kiwala S, McMichael J, et al (2020). Full citation and original link in Sources, entry 36.

Study 4 Identify an actionable residual disease window

Put serial ctDNA, standard treatment, vaccine manufacturing and administration dates, imaging, and clinical events on one timeline. Study false-negative tests, differences in DNA shedding across lesions, and whether time between molecular detection and visible metastasis leaves a real opportunity for intervention.

Controls should have comparable testing frequency and standard treatment. More intensive testing in an intervention group must not make earlier detection appear to be worse efficacy. ctDNA clearance and immune responses can be exploratory intermediate measures; clinical claims still require appropriate designs and long-term outcomes.[7,8]

Study 5 Link product and safety data to mechanism

Development must connect the intended design to the delivered product: candidate sequence versions, delivery or cell platform, batch identity, purity and potency assays, storage, transport, production time, and failures. Platforms need different quality and nonclinical safety evidence. A database cannot supply universal product-release thresholds.[9]

Normal cells and tissues, patient-derived models, and appropriate immune experiments can support staged candidate validation. Each model should have an explicit account of what it can and cannot answer. HCMI provides genomically linked experimental models, not a replacement for a complete immune system or human safety studies.[31]

Study 6 Establish clinical benefit beyond immunogenicity

Estimating benefit added to current standard treatment requires appropriate controls, prespecified statistical plans, complete enrollment denominators, clinical endpoints, and sufficient follow-up. Retain manufacturing failures, rapid progression before vaccination, withdrawals, and loss to follow-up. Do not analyze only the patients who completed every vaccination.

Incomplete public immune-correlate data are an important gap. Aggregate negative clinical results alone cannot distinguish poor antigen selection, inadequate immune responses, microenvironmental obstruction, or an unsuitable clinical setting. The incompletely public correlative endpoints in the FR-alpha registration illustrate why controlled sharing of mechanisms and outcomes should be planned from the start of a trial.[16]

Sample size must follow the specific hypothesis, feasibility, event rate, and statistical power. This report does not invent a universal number of patients sufficient to discover a vaccine.

6 Data development priorities for Tesserie

What participant uploads can contribute

Original imaging, reports, pathology, medication histories, and follow-up can help recover fragmented care histories and establish verifiable timelines for future collaborative research. This is valuable work, but should be described as participant-authorized longitudinal records and research linkage. Uploading an ultrasound PDF does not enable development of a personalized cancer vaccine.

Data source	Realistic contribution to vaccine research	Capability it cannot establish
Participant PDFs and reports	Identify pathological subtype, care dates, testing institutions, and retrievable original records	A PDF is not original DICOM, a whole-slide image, or a sequencing file
Original DICOM and digital pathology	Quantify lesions, treatment response, and pathological features; connect later events	Cannot alone establish HLA antigens or T-cell recognition
EHR and longitudinal participant follow-up	Reconstruct treatment, recurrence, concomitant medications, and quality of life, retaining sources and uncertainty	Self-reported outcomes are not automatically clinically adjudicated research endpoints
Wearable and symptom data	Support adherence, physical function, symptom, and quality-of-life research	Usually not the highest priority for antigen discovery; cannot establish tumor clearance
Hospital and biospecimen partnerships	Obtain appropriately authorized tissues, paired blood, sequencing, immune assays, and clinical verification	Permission to upload files does not authorize new experiments, commercial model training, or all institutional data use

A minimum linked research structure

A participant table would hold research identifiers and consent versions. A clinical-course table would record diagnosis and treatment events. A lesion table would distinguish primary, residual, recurrent, and metastatic disease. A specimen table would retain collection dates, sites, and processing. Assay tables would preserve platforms, original file locations, quality control, and analysis versions. Antigen and immune-assay tables would hold specific evidence. An outcome table would preserve event definitions, dates, sources, and verification status.

Times should be anchored relative to diagnosis, surgery, standard treatment, and vaccination. Every file should connect to its source and relevant patient, specimen, or lesion. De-identified sharing should preserve time intervals and separate direct identity from research identifiers. Missingness should distinguish not collected, unavailable, experimental failure, failed quality control, and not applicable.

This is a proposed research architecture, not a description of current website capabilities. No real patient records were collected, uploaded, or processed in preparing this review.

Sequence of development

The first stage would establish authorization, source verification, and structured clinical timelines to demonstrate reliable linkage of records. A second stage would work with hospitals or research institutions to connect tissue, paired blood, and raw experimental data in appropriately ethics-approved cohorts. A third would establish prospective paired validation around one specific vaccine question, progressively including antigen evidence, functional immunity, and clinical outcomes.

If resources are limited, we would prioritize a smaller paired cohort with complete treatment timelines, original molecular and immune assays, and sustained follow-up over a large collection of files lacking paired biospecimens and outcomes. This is a product and research strategy judgment based on the evidence above, not a validated business model or sample-size conclusion.

Permissions are part of research feasibility

NIH controlled data commonly require institutional and project applications, review of intended uses, and data-use commitments. A public paper or visible accession does not authorize use of controlled raw data.[37] For each resource, separately verify download access, research and commercial uses, model training, redistribution restrictions, and material-transfer terms. Do not attempt to re-identify participants.

Particular constraints include Duke MRI's noncommercial license, academic-use restrictions on TONIC-related sequence and response data, and the conditional sharing plan for NCT04674306.[15,29,33] Check current terms before collaboration. This report does not replace legal review or data-use approval.

7 Practical research starting points

Begin with publicly accessible data directly relevant to the question. Breast cancer PXD034818 and its transcript data can establish candidate presentation evidence. HLA Ligand Atlas, the normal breast atlas, GTEx, and Human Protein Atlas can supply normal-tissue context. CEDAR and TESLA can inform the organization of functional labels and negative controls. Do not treat samples from these different resources as the same patients.

Next, apply for actual paired breast cancer vaccine data, such as dbGaP phs002787.v2.p1, and confirm which patient, antigen, immune-response, and outcome links are available. Missing links should become explicit research limitations or collaboration requirements, not be filled with results from another cancer or cohort.

Finally, turn unanswered questions into plans for collaborative data collection. Which patients' antigens were recognized without clinical benefit? Who could not receive vaccination because of manufacturing or disease progression? Were the targets still present at recurrence? Do the patterns hold across subtypes and HLA groups? Real paired data, including failures, that can answer these questions would make a valuable contribution to breast cancer vaccine development.

37 National Institutes of Health (2026). Full citation and original link in Sources, entry 37.

Sources

1. Sahin U, Schmidt M, Derhovanessian E, et al. Individualized mRNA vaccines evoke durable T cell immunity in adjuvant TNBC. Nature. 2026;651:1088-1096. Published 18 February 2026. Original source

2. Zhang X, Goedegebuure SP, Chen MY, et al. Neoantigen DNA vaccines are safe, feasible, and induce neoantigen-specific immune responses in triple-negative breast cancer patients. Genome Medicine. 2024. DOI 10.1186/s13073-024-01388-3. Original source

3. US Food and Drug Administration. Clinical Considerations for Therapeutic Cancer Vaccines. Final guidance, October 2011. Scope excludes primary prevention in people without a cancer history. Original source

4. Kina E, Laverdure JP, Durette C, et al. Breast cancer immunopeptidomes contain numerous shared tumor antigens. Journal of Clinical Investigation. 2024;134(1):e166740. First published 31 October 2023. Includes accessions PXD034818, PRJNA852282 and GSE206838. Original source

5. Marcu A, Bichmann L, Kuchenbecker L, et al. HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. Journal for ImmunoTherapy of Cancer. 2021;9:e002071. Original source

6. Wells DK, van Buuren MM, Dang KK, et al. Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. Cell. 2020;183:818-834.e13. TESLA consortium. Original source

7. US Food and Drug Administration. Use of Circulating Tumor DNA for Curative-Intent Solid Tumor Drug Development. Final guidance, November 2024. Original source

8. Turner NC, Swift C, Jenkins B, et al. Results of the c-TRAK TN trial: a clinical trial utilising ctDNA mutation tracking to detect molecular residual disease and trigger intervention in patients with moderate- and high-risk early-stage triple-negative breast cancer. Annals of Oncology. 2023;34:200-211. Original source

9. US Food and Drug Administration. Potency Assessment of Active Immunotherapy Products. Draft guidance, August 2026; not for implementation. Original source

10. IPD-IMGT/HLA Database. Official HLA allele nomenclature and reference sequences. Release 3.65, July 2026; accessed 8 September 2026. Original source

11. Allele Frequency Net Database. Population allele and haplotype frequency resources and quality classification. Accessed 8 September 2026. Original source

12. Nishida J, Seehawer M, Rojas Jimenez E, et al. Breast cancer prevention by prophylactic Lalba mRNA-LNP vaccination. Proceedings of the National Academy of Sciences. 2026. DOI 10.1073/pnas.2621855123. Animal study. Original source

13. Virassamy B, Caramia F, Savas P, et al. Parity and lactation induce T-cell-mediated breast cancer protection. Nature. 2026;649:449-459. First online 20 October 2025. Original source

14. Storrs E, Mo CK, Chou WH, et al. Spatial Mapping of the Precancer-to-Cancer Transition in Breast and Prostate. Cancer Discovery. Published online 17 April 2026. DOI 10.1158/2159-8290.CD-26-0012. Original source

15. ClinicalTrials.gov. NCT04674306, alpha-lactalbumin breast cancer vaccine study. Registry and individual participant data sharing plan; updated 17 July 2026. Original source

16. ClinicalTrials.gov. NCT03012100, TPIV200 folate receptor alpha vaccine study. Results and correlative outcomes; updated 30 June 2026. Original source

17. NCBI dbGaP. Personalized Breast Cancer Vaccines Based on Genome Sequencing. phs002787.v2.p1. Current study metadata and authorized-access entry; accessed 8 September 2026. Original source

18. Koşaloğlu-Yalçın Z, Carri I, Marrama D, et al. The Cancer Epitope Database and Analysis Resource (CEDAR): current capabilities and future directions. Frontiers in Oncology. Published 19 August 2026. Provider-authored resource review. Original source

19. Cancer Epitope Database and Analysis Resource. Database export files and license. Download listing updated 1 September 2026; accessed 8 September 2026. Original source

20. National Cancer Institute Genomic Data Commons. Data Access Processes and Tools. Open versus controlled data access; accessed 8 September 2026. Original source

21. Krug K, Jaehnig EJ, Satpathy S, et al. Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. Cell. 2020;183:1436-1456.e31. Includes PDC000120. Original source

22. cBioPortal. Breast Cancer METABRIC study. Study and modality-specific coverage; accessed 8 September 2026. Original source

23. Garcia-Recio S, Hinoue T, Wheeler GL, et al. Multiomics in primary and metastatic breast tumors from the AURORA US network finds microenvironment and epigenetic drivers of metastasis. Nature Cancer. 2023;4:128-147. Published online 30 December 2022. Original source

24. Wu SZ, Al-Eryani G, Roden DL, et al. A single-cell and spatially resolved atlas of human breast cancers. Nature Genetics. 2021;53:1334-1347. Original source

25. Kumar T, Nee K, Wei R, et al. A spatially resolved single-cell genomic atlas of the adult human breast. Nature. 2023;620:181-191. Original source

26. GTEx Consortium. GTEx adult data downloads and access overview. Accessed 8 September 2026. Original source

27. Human Protein Atlas. Data download and license information. Version 25.1; accessed 8 September 2026. Original source

28. Human Tumor Atlas Network. Sharing data from the Human Tumor Atlas Network through standards, infrastructure and community engagement. Nature Methods. 2025. DOI 10.1038/s41592-025-02643-0. Original source

29. Greenwald NF, Nederlof I, et al. Temporal and spatial composition of the tumor microenvironment predicts response to immune checkpoint inhibition in metastatic TNBC. Nature Cancer. 2026;7:435-450. Published 18 February 2026. See data availability for commercial-use restrictions. Original source

30. Ranek JS, Greenwald NF, Goldston M, et al. The automated computational workflow QUICHE reveals structural definitions of antitumor responses in triple-negative breast cancer. Nature Cancer. Published 18 February 2026. Original source

31. El Harouni D, Al-Jazrawe M, Choi S, et al. A compendium of next-generation patient-derived models for diverse cancers. Nature. Published 5 August 2026. Original source

32. The Cancer Imaging Archive. I-SPY2 Imaging Collection. Data descriptions, cohort overlap and CC BY 4.0 license; accessed 8 September 2026. Original source

33. The Cancer Imaging Archive. Duke Breast Cancer MRI Collection. Data descriptions and CC BY-NC 4.0 license; accessed 8 September 2026. Original source

34. ImmPort. Shared data and documentation. Accessed 8 September 2026. Original source

35. Deutsch EW, Kok LW, Mudge JM, et al. Expanding the human proteome with microproteins and peptideins. Nature. 2026;654:813-825. Published 6 May 2026. Original source

36. Hundal J, Kiwala S, McMichael J, et al. pVACtools: A Computational Toolkit to Identify and Visualize Cancer Neoantigens. Cancer Immunology Research. 2020;8:409-420. Original source

37. National Institutes of Health. Accessing Scientific Data. Data access and use requirements; accessed 8 September 2026. Original source