International Journal of Engineering Research and Modern Education (IJERME)
Impact Factor: 7.018, ISSN (Online): 2455 - 4200
(www.rdmodernresearch.com) Volume 8, Issue 2, 2023

NEOANTIGEN PRESENTATION FEATURES AND IMMUNOGENICITY FOR PERSONALIZED CANCER VACCINE TARGET SELECTION ACROSS TUMORS DURING 2010-2022

Kannan Vishwanath* & Lloyd Zulu**
* Garwood Innovation Fellow, Garwood Center for Corporate Innovation, Haas School of Business, Berkeley
** World Christian University, Lusaka, Zambia

Cite This Article: Kannan Vishwanath & Lloyd Zulu, "Neoantigen Presentation Features and Immunogenicity for Personalized Cancer Vaccine Target Selection Across Tumors During 2010-2022", International Journal of Engineering Research and Modern Education, Volume 8, Issue 2, July - December, Page Number 70-80, 2023.
Copy Right: © IJERME, 2023 (All Rights Reserved). This is an Open Access Article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract:

This analysis examines whether molecular profiles used in personalized cancer vaccine selection remain consistent across patients, addressing the transferability of pooled candidate evidence. A published cross-sectional dataset provides 404 complete biological sample records from six patients, including 29 records with experimentally validated immunogenicity. Descriptive comparisons, patient specific medians, Spearman correlations, and exploratory sensitivity checks assess three presentation features. Positive records show median predicted affinity of 17.50 versus 74.30 nanomolar, abundance of 40.44 versus 11.38 transcripts per million, and stability of 3.28 versus 2.10 hours. These patterns are compatible with complementary constraints involving binding, transcript availability, and complex persistence. However, within patient directions vary, with stability showing the pooled ordering in five patients and the other features in four. Equal patient weighting and removal of repeated complexes preserve pooled directions while changing magnitudes. The contribution distinguishes aggregate molecular separation from patient specific consistency, qualifying presentation-based selection frameworks. Research teams should report patient variation and coverage losses alongside pooled comparisons when prioritizing candidates for experimental assessment.

Key Words: Neoantigen Immunogenicity; Personalized Cancer Vaccines; Antigen Presentation; Patient Heterogeneity; Descriptive Analysis

1. Introduction:

Personalized cancer vaccines require a distinction between mutations that generate candidate targets and peptides that the immune system recognizes. Selecting a peptide with favorable molecular characteristics does not establish its immunogenicity. This distinction matters because target selection determines which candidates advance from computational screening to experimental assessment and vaccine construction. Clinical research has demonstrated the feasibility of inducing responses against selected personal neoantigens, but responses vary across candidates and patients (Ott et al., 2017). Understanding that variation requires attention to both molecular features and the composition of the evidence used to evaluate them.

Antigen presentation provides the biological basis for examining predicted binding affinity, tumor abundance, and predicted binding stability. Affinity describes the strength of peptide binding to a compatible human leukocyte antigen molecule. Abundance represents the availability of transcript material associated with the candidate, while stability describes the predicted persistence of the resulting complex. These properties address different constraints between mutation expression and immune recognition. Expression information can improve assessment of antigen presentation, but transcript abundance does not directly measure the density of displayed peptide complexes (Abelin et al., 2017). Likewise, predicted binding strength and complex persistence provide related rather than interchangeable information (Rasmussen et al., 2016).

The distinction between presentation and recognition also limits what favorable molecular values imply. A candidate can bind strongly or remain associated with an antigen molecule without eliciting a detectable response in the assessed immune repertoire. Strønen et al. (2016) demonstrated that donor T cells could recognize neoantigens neglected by patients' own T cell populations. Such evidence places immune recognition beyond a simple ranking of presentation characteristics. It also supports interpreting experimentally validated immunogenicity as a distinct outcome rather than treating predicted binding as its substitute.

The closest empirical foundation is the Tumor Neoantigen Selection Alliance investigation. Wells et al. (2020) compared candidate predictions from multiple teams and experimentally assessed their immunogenicity using patient matched samples. Their analysis identified binding affinity, tumor abundance, and binding stability as relevant presentation features. This evidence already establishes the importance of examining these characteristics together. A defensible extension therefore requires more than reproducing pooled differences between positive and negative candidates.

The question addressed here concerns how those pooled differences relate to variation within patients and to analytical sample construction. Patients contribute unequal numbers of candidates, transcript abundance is unavailable for some records, and a small number of peptide complexes appear in different biological samples. Consequently, a pooled contrast can reflect both molecular separation and the relative representation of patients. The reviewed evidence does not establish that the magnitude or direction of every pooled difference will remain uniform within each patient. Examining this distinction qualifies how aggregate candidate characteristics inform personalized selection.

This study aims to characterize differences in predicted binding affinity, tumor abundance, and predicted binding stability between records with and without experimentally validated immunogenicity, and to assess their consistency across patients and alternative descriptive summaries. The central research question is whether the pooled presentation profile of validated candidates persists when patient variation and repeated biological sample records are considered. The objectives are to compare each presentation feature by validation status, examine the corresponding patient specific contrasts, and evaluate sensitivity to equal patient weighting and counting repeated complexes once. These objectives concern descriptive relationships rather than causal effects or clinical vaccine efficacy.

The analysis uses the published supplementary dataset associated with Wells et al. (2020). From 608 source records, 404 contain all three presentation features and the validation outcome. The retained sample includes 29 positive records from six patients and represents 401 distinct patient and peptide HLA combinations. The observational structure is a cross section of assessed candidates rather than a longitudinal patient panel. The analysis compares distributions, patient specific medians, and descriptive correlations. Its methodological value lies in making the relationship between pooled patterns and patient composition explicit.

The findings show that positive records have stronger median predicted binding, greater median abundance, and longer median predicted stability. These directions persist when patient contributions receive equal weight and when repeated complexes are counted once. However, individual patients do not uniformly reproduce the pooled ordering, and the affinity contrast becomes smaller under equal patient weighting. The contribution is therefore an empirical qualification of aggregate presentation profiles: directional persistence across descriptive checks can coexist with variation in magnitude and patient specific exceptions.

This distinction has relevance for research teams evaluating personalized vaccine targets across clinical settings. It supports reporting patient specific contrasts and coverage losses alongside pooled candidate summaries, rather than assuming that an aggregate ordering applies equally to every patient. Personalized vaccine development involves interconnected decisions about candidate identification, immune assessment, and individualized production (Blass & Ott, 2021). The present evidence informs the evaluation of the candidate selection stage. Its transferable contribution is the analytical distinction between pooled separation and patient level consistency, while its empirical conclusions remain bounded by the assessed candidates and patients.

2. Theory Conceptual Framework and Research Questions Development:

Personalized cancer vaccine target selection rests on a sequence of biological constraints. A tumor mutation must generate an expressed peptide, the peptide must bind to a compatible human leukocyte antigen molecule, and the resulting complex must remain available for recognition by T cells. Schumacher and Schreiber in 2015 explain that tumor mutations provide potential immune targets, but only a subset generates effective immune recognition. This analysis applies that antigen presentation logic to three observed features: predicted binding affinity, tumor abundance, and predicted binding stability. The outcome is experimentally validated neoantigen immunogenicity, coded as one for a validated response and zero otherwise.

These features represent distinct constraints within an interdependent process. Binding affinity concerns the strength of peptide attachment, abundance concerns the availability of its source material, and stability concerns the persistence of the resulting complex. None alone establishes immune recognition. Yadav et al. in 2014 demonstrated the value of integrating tumor sequencing, expression information, and direct peptide identification when selecting immunogenic tumor mutations. Their findings support a framework in which presentation features guide target selection while experimental immune recognition remains the outcome requiring verification.

Predicted binding affinity measures the strength of the interaction between a peptide and its restricting antigen molecule. This analysis uses the predicted affinity values recorded in nanomolar units. Lower values indicate stronger predicted binding. Stronger binding can increase the likelihood that a peptide forms a complex capable of reaching the cell surface, providing a biological basis for expecting lower affinity values among immunogenic peptides.

Binding nevertheless provides an incomplete explanation. A peptide can bind strongly without encountering a responsive T cell population. Prediction error can also separate estimated binding strength from biological presentation. The theoretical expectation therefore concerns an association with experimentally validated immunogenicity rather than a deterministic relationship.

Wells et al. in 2020 identified binding affinity among the presentation features associated with neoantigen immunogenicity in the source dataset. Their finding provides empirical motivation, but the present

analysis distinguishes the recorded predicted affinity measure from experimentally measured affinity. It also treats analysis of these records as exploratory reassessment of an existing dataset rather than independent confirmation.

RQ1: How does predicted binding affinity differ between experimentally validated immunogenic and nonimmunogenic peptides?

Tumor abundance addresses the availability of material from which a candidate peptide can be generated. This analysis uses the recorded abundance measure in transcripts per million. Greater abundance can increase the supply of source material available for peptide processing and presentation. The theoretical expectation is therefore that immunogenic peptides will tend to have greater tumor abundance.

Abundance differs from affinity because greater supply does not ensure successful binding or recognition. Transcript abundance also does not directly measure the number of peptide complexes displayed at the cell surface. Processing efficiency and antigen binding can constrain presentation even when the originating transcript is abundant.

Abelin et al. in 2017 showed that expression information contributes to prediction of antigen presentation alongside peptide binding characteristics. This evidence supports treating abundance as a separate observed predictor while retaining the distinction between transcript availability, peptide presentation, and experimentally validated immunogenicity.

RQ2: How does tumor abundance differ between experimentally validated immunogenic and nonimmunogenic peptides?

Predicted binding stability concerns the duration of the peptide interaction with its antigen molecule. This analysis uses the recorded stability estimates in hours. Longer persistence can extend the opportunity for T cells to encounter and recognize a displayed peptide. Stability therefore introduces a temporal constraint that differs conceptually from binding strength.

Harndahl et al. in 2012 found that complex stability predicted cytotoxic T cell immunogenicity better than affinity in their experimental setting. This supports the proposed persistence mechanism. It does not establish that stability will outperform affinity in every tumor dataset, particularly when both quantities are predicted rather than directly measured.

Rasmussen et al. in 2016 showed that combining predicted stability with predicted affinity improved epitope prediction. Their findings support examining stability alongside affinity while recognizing that the two features can share information. The expected pattern is greater stability among immunogenic peptides, subject to the requirement for a compatible recognition response.

RQ3: How does predicted binding stability differ between experimentally validated immunogenic and nonimmunogenic peptides?

The conceptual framework connects three observed presentation features with one observed binary outcome. These connections represent theoretical associations, not identified causal effects. The analysis uses 404 complete peptide records, including 29 validated immunogenic records, nested within six patients. Comparisons therefore retain patient identity and examine whether patterns remain consistent within patients. The framework introduces no latent construct, composite index, mediator, or moderator. Its research questions permit positive, negative, or inconclusive findings without changing the proposed biological interpretation after observing results.

3. Data:

This section establishes data provenance, documents sample construction, and defines the observed measures. It reports completed quality checks and descriptive comparisons while distinguishing patients, peptide complexes, and biological samples.

3.1 Data Sources, Population, and Sample Construction:

The analysis uses the Tumor Neoantigen Selection Alliance dataset published by Wells et al. (2020). The source is supplementary Table S4, worksheet master-bindings-selected, available through the article supplement and the Caltech institutional repository. The associated repository identifier is syn21048999. The supplied workbook preserves the source records, analytical sample, and extraction metadata. No numbered revision or explicit reuse licence is recorded in the workbook.

The source cohort comprises six patients, including three melanoma cases and three non small cell lung cancer cases. Candidates were selected from ranked computational predictions and assessed experimentally rather than sampled randomly from all tumor mutations (Wells et al., 2020). The accessible sampling frame therefore represents selected candidate neoantigens rather than the wider population of cancer patients or tumor mutations.

The analytical unit is a patient, peptide HLA complex, and biological sample combination. The dataset forms a cross section of experimentally assessed candidates. It does not contain repeated clinical observation periods. Sampling dates and a common clinical observation interval are unavailable in the analytical table, so publication timing does not define a longitudinal observation period.

The published table contains 608 records from six patients, representing 605 distinct patient and peptide HLA combinations. Three combinations occur in both blood and tumor samples. The implemented inclusion rule requires observed predicted affinity, tumor abundance, predicted stability, and validation status. This rule excludes 204 records with unavailable tumor abundance and retains 404 records representing 401 distinct patient and peptide HLA combinations.

Table 1: Sample construction

Stage	Patients Represented	Records	Inclusion or Exclusion Rule
Published experimental frame	6	608	All records in supplementary Table S4
Excluded records	6	204	Tumor abundance unavailable
Complete analytical sample	6	404	All four analytical variables observed
Validated immunogenic records	6	29	Validation status coded one
Records without validated immunogenicity	6	375	Validation status coded zero

The inclusion rule was documented during dataset construction before the descriptive comparisons. No additional records were removed through merging, transformation, lag construction, or estimator requirements. Retained observations range from 19 to 90 per patient.

3.2 Data Integration, Cleaning, and Quality Assurance:

Preparation uses one published source table without external merging. The analytical table retains patient identifiers, peptide HLA identifiers, peptide sequences, HLA alleles, biological sample types, and source row references. These identifiers permit reconciliation with the original supplement. All retained predictor values match their corresponding source values.

Checks identified no identical full records and no duplicate combinations of patient identifier, peptide HLA identifier, and sample type. Three patient and peptide HLA combinations appear in both blood and tumor samples. Both records were retained because they represent different biological samples. These combinations belong to patient 12 and have negative validation outcomes in both samples.

Tumor abundance is unavailable for 204 of 608 source records, equivalent to 33.55 percent. Predicted affinity, predicted stability, and validation status are observed for every source record. Missing abundance affects all six patients, with unequal losses across patients.

Table 2: Patient coverage and missing abundance

Patient Identifier	Source Records	Missing Abundance	Retained Records	Retained Positive Outcomes
1	97	27	70	8
2	108	21	87	4
3	97	22	75	10
10	73	54	19	2
12	89	26	63	2
16	144	54	90	3
Total	608	204	404	29

Four recorded zero abundance values were retained as observed source values. Range checks confirmed positive predicted affinity and stability, nonnegative abundance, and binary outcome coding. Extreme observations remained unchanged. No trimming, winsorization, logarithmic transformation, or standardization was applied.

Eight positive outcomes occur among excluded records. Positive outcomes account for 7.18 percent of retained records and 3.92 percent of excluded records. Complete record selection therefore changes outcome composition and patient representation. These comparisons document selection without identifying its cause.

3.3 Variable Measurement, Construction, and Validation:

The analysis represents neoantigen presentation through three observed features rather than a latent construct or composite index. Predicted affinity measures peptide HLA binding strength. Tumor abundance measures transcript availability, while predicted stability measures complex persistence. Expression information contributes to antigen presentation assessment but does not directly measure the quantity of peptide displayed at the cell surface (Abelin et al., 2017).

Predicted stability is retained in hours. The source investigators calculated this feature using NetMHCstabpan, whose methodological basis links predicted complex persistence with T cell immunogenicity

(Rasmussen et al., 2016). These values are inherited predictions rather than laboratory measurements newly generated for this analysis.

Table 3: Variable dictionary

Variable	Analytical Role	Definition and Source Identifier	Unit and Frequency	Transformation and Interpretation	Missing Value Rule	Coverage and Limitation
Predicted binding affinity	Presentation feature	Source predicted affinity, NETMHC_PAN_BINDING_AFFINITY	Nanomolar; one value per record	Untransformed; lower values indicate stronger predicted binding	Required for inclusion	608 source and 404 retained records; predicted rather than measured affinity
Tumor abundance	Presentation feature	Source transcript abundance, TUMOR_ABUNDANCE	Transcripts per million; one value per record	Untransformed; higher values indicate greater abundance	Exclude unavailable values; retain observed zeros	404 observed records; indirect measure of peptide availability
Predicted binding stability	Presentation feature	Estimated complex persistence, BINDING_STABILITY	Hours; one value per record	Untransformed; higher values indicate longer persistence	Required for inclusion	608 source and 404 retained records; predicted duration
Validated immunogenicity	Outcome	Source VALIDATED status recoded to one or zero	Binary; one status per record	One indicates validated immunogenicity	Required for inclusion	37 positive source records and 29 retained positive records
Patient identifier	Grouping identifier	Original PATIENT_ID retained unchanged	Categorical; record level	No numerical interpretation	Required for identification	Six patients with unequal candidate coverage
Biological sample	Record identifier	Original TISSUE_TYPE classification	Categorical; record level	Blood or tumor sample category	No missing source values	Distinguishes repeated complexes across biological samples

The outcome preserves the experimental classification supplied by Wells et al. (2020). The source investigators assessed peptide specific T cell recognition; the present analysis checks the coding without repeating those experiments. A zero indicates that immunogenicity was not validated under the recorded assessment.

Affinity and stability remain separate because binding strength and complex persistence represent different molecular properties. Experimental evidence supports distinguishing their contributions to immunogenicity assessment (Harndahl et al., 2012). Their conceptual distinction does not establish statistical independence in this sample.

Table 4: Descriptive statistics for retained records

Variable	Valid Records	Mean	Standard Deviation	Median	Minimum	Maximum
Predicted affinity in nanomolar	404	231.84	707.40	64.80	3.00	7438.80
Tumor abundance in transcripts per million	404	33.20	57.00	12.10	0.00	460.08
Predicted stability in hours	404	3.85	5.57	2.16	0.12	55.36
Validated immunogenicity	404	0.0718	0.2584	0	0	1

Standard deviations use the sample definition. Predictor means exceed their medians, and observed ranges show substantial dispersion. Reporting both means and medians preserves this distributional information without imposing additional transformations.

3.4 Analytical Framework, Identification, and Reproducibility:

The implemented analysis is descriptive. It examines sample composition, variable distributions, and differences between records with and without validated immunogenicity. Reported quantities comprise counts, proportions, means, medians, standard deviations, and ranges. No regression estimates, significance tests, causal effects, or predictive performance measures are reported.

Median predicted affinity is 17.50 nanomolar among positive records and 74.30 among negative records. Median abundance is 40.44 transcripts per million among positive records and 11.38 among negative

records. Median predicted stability is 3.28 hours among positive records and 2.10 among negative records. These differences describe the retained sample without establishing independent predictor effects.

The interpretation remains confined to experimental immunogenicity. Antigen generation, presentation, and immune recognition form related biological processes, but presentation characteristics alone do not establish therapeutic benefit (Schumacher & Schreiber, 2015).

Multiple candidates belong to each patient, and three patient and peptide combinations occur in two biological samples. Pooled summaries therefore describe records rather than independent patients. Unequal patient coverage and missing abundance further constrain generalization beyond the assessed candidate set.

Reproducibility rests on the preserved source worksheet, analytical worksheet, original identifiers, and source row references. Retained values were reconciled with the source, and outcome coding was checked against the original validation status. All predictors use the same 404 record sample and retain their original units, supporting consistent descriptive comparison.

4. Methods:

4.1 Research Design and Analytical Scope:

The analysis uses a descriptive observational design to examine differences in three presentation features between candidates with and without experimentally validated immunogenicity. The unit of analysis is a patient, peptide HLA complex, and biological sample combination. The outcome separates records with validated immunogenicity from records without validated immunogenicity under the source assessment. The comparison concerns assessed candidates rather than treatment groups or clinical outcomes.

The analytical questions address pooled differences, consistency within patients, and sensitivity to patient representation and repeated complexes. Variation arises across candidates within and between six patients. Multiple records belong to each patient, and some complexes occur in different biological samples. These relationships are retained when interpreting the pooled summaries.

The source investigation assessed candidates selected from computational predictions rather than a probability sample of all tumor mutations (Wells et al., 2020). Accordingly, the analytical target is the distribution of presentation characteristics within the accessible candidate set. The design supports descriptive comparisons without identifying causal effects or independent predictor contributions.

4.2 Sample and Measurement Alignment:

The analysis follows the eligibility rules documented in Table 1 and the measurement definitions in Table 3. The analytical sample contains 404 records with observed predicted affinity, tumor abundance, predicted stability, and validation status. All descriptive comparisons use this common sample unless an alternative sample is explicitly identified. The retained records represent six patients and 401 distinct patient and peptide HLA combinations.

Predicted affinity remains in nanomolar units, abundance in transcripts per million, and predicted stability in hours. Lower affinity values indicate stronger predicted binding, whereas higher abundance and stability values indicate greater transcript availability and longer predicted persistence. Validation status is recoded from the source Boolean classification to one for validated immunogenicity and zero otherwise.

No normalization, logarithmic transformation, composite construction, or threshold based classification is applied. The three presentation features remain separate measures. Patient identifiers define comparison groups rather than numerical predictors, while sample type distinguishes blood and tumor records. The source investigators generated the experimental outcome classification; this analysis preserves that classification rather than repeating its laboratory validation.

Missing abundance determines the complete record restriction. Recorded zeros remain observed values and are not treated as missing. Predicted rather than measured affinity is used consistently across retained records, avoiding substitution between measurement types.

4.3 Descriptive Procedures and Patient Comparisons:

Calculations were performed using Python 3.12.14, pandas 2.2.3, and NumPy 2.3.5. Processing began by reading the source and analytical worksheets, checking identifiers, and reconciling retained predictor values with their corresponding source records. Missingness was counted for each analytical variable and summarized by patient. Outcome proportions were calculated separately for retained and excluded records.

For each presentation feature, the analysis calculated the valid count, arithmetic mean, sample standard deviation, median, minimum, and maximum. Sample standard deviations use a denominator equal to the number of observations minus one. The binary outcome was summarized through counts and proportions. Predictor summaries were then calculated separately for positive and negative outcome groups, producing the comparisons reported in Table 5.

Median contrasts were calculated by subtracting the negative group median from the positive group median. Relative contrasts divided this difference by the negative group median and multiplied the result by 100. Ratios divided the positive group median by the negative group median. Calculations used unrounded values; rounding was applied only for presentation. These quantities describe differences between group summaries rather than changes attributable to an intervention.

Patient specific comparisons were constructed by grouping records jointly by patient identifier and validation status, then calculating the median of each presentation feature within each group. Table 6 reports these medians. Directional consistency was assessed by comparing positive and negative group medians within each patient. All six patients contributed records to both outcome groups.

Pairwise relationships among the three presentation features were summarized using Spearman rank correlations on the same 404 records. Tied observations received average ranks. These coefficients describe monotonic correspondence between recorded features without adjusting for patient membership. They are interpreted as descriptive measures rather than estimates based on independent patient observations.

4.4 Quality Checks, Sensitivity, and Reproducibility:

Quality checks examined missing values, unique record identifiers, identical full records, and repeated combinations of patient identifier, peptide HLA identifier, and sample type. Range checks assessed whether affinity and stability were positive, abundance was nonnegative, and outcome values were binary. Observed extreme values were retained without trimming or winsorization.

Two exploratory sensitivity checks addressed specific features of the sample. The first examined unequal patient representation. Within each outcome group, the analysis averaged the six patient specific medians, giving each patient equal weight. These summaries were compared with the pooled medians by direction and magnitude. This procedure is an average of patient medians, not a weighted median of individual records.

The second check examined repeated patient and peptide HLA combinations across biological samples. One record per combination was retained in the original worksheet order, reducing the sample to 401 distinct combinations. Before this restriction, the analysis verified that each repeated combination had identical values for all three presentation features and the outcome. The choice of retained biological sample therefore did not change those analytical values. Outcome group medians were recalculated on the restricted sample.

These sensitivity checks were introduced as exploratory assessments after the pooled comparisons. Equal patient weighting changes the contribution of patients without removing records from the underlying calculations, whereas counting each complex once changes the analytical sample. Their results are interpreted accordingly. Neither procedure establishes independence or removes selection arising from missing abundance.

Replication relies on the preserved source worksheet, analytical worksheet, patient identifiers, biological sample labels, and source row references. All reported comparisons follow the same outcome coding and measurement units. The methodological contribution is a transparent assessment of pooled patterns, patient variation, and sample sensitivity using established descriptive procedures.

5. Findings:

This analysis examines how predicted binding affinity, tumor abundance, and predicted binding stability differ by experimentally validated immunogenicity. The findings address the three research questions through descriptive comparisons of 404 complete records. Patient level comparisons and sensitivity checks assess whether pooled patterns depend on unequal coverage or repeated peptide complexes.

5.1 Sample Composition and Distributional Characteristics:

The analytical sample contains 404 biological sample records from six patients, representing 401 distinct patient and peptide HLA combinations. Of these records, 29 show validated immunogenicity and 375 do not, giving an observed positive proportion of 7.18 percent. The outcome retains the experimental classification supplied by Wells et al. (2020), while the comparisons reported here are calculated from the supplied analytical dataset.

Complete record selection excludes 204 observations with missing tumor abundance, equivalent to 33.55 percent of the source records. Eight excluded records have positive outcomes. The positive proportion is 3.92 percent among excluded records, compared with 7.18 percent among retained records. Consequently, the analytical sample contains a greater proportion of positive outcomes than the excluded sample.

Patient contributions range from 19 to 90 retained records. Positive outcomes range from two to ten per patient. These differences matter because pooled summaries assign greater weight to patients contributing more candidate records. The sample therefore supports comparison of assessed candidates rather than estimation of immunogenicity prevalence across cancer patients.

Table 5: Presentation features by validated immunogenicity

Feature	Outcome Group	Records	Mean	Standard Deviation	Median	Minimum	Maximum
Predicted affinity in nanomolar	Positive	29	57.32	97.79	17.50	4.20	415.50
Predicted affinity in nanomolar	Negative	375	245.34	732.09	74.30	3.00	7438.80
Tumor abundance in transcripts per million	Positive	29	39.87	32.75	40.44	0.90	110.06

Feature	Outcome Group	Records	Mean	Standard Deviation	Median	Minimum	Maximum
Tumor abundance in transcripts per million	Negative	375	32.68	58.46	11.38	0.00	460.08
Predicted stability in hours	Positive	29	6.05	9.38	3.28	0.14	49.34
Predicted stability in hours	Negative	375	3.68	5.15	2.10	0.12	55.36

The distributions overlap across outcome groups. Negative records include affinity values indicating strong predicted binding, high abundance values, and long predicted stability. Positive records also include comparatively low abundance and short predicted stability. None of the three features individually separates all positive from negative records.

5.2 Predicted Binding Affinity and Immunogenicity:

The first research question concerns differences in predicted binding affinity. Positive records have a median affinity of 17.50 nanomolar, compared with 74.30 among negative records. The median is therefore 56.80 nanomolar lower, equivalent to a 76.45 percent reduction relative to the negative group. Because lower affinity values indicate stronger predicted binding, this comparison identifies stronger median predicted binding among positive records.

The mean comparison points in the same direction. Mean affinity is 57.32 nanomolar among positive records and 245.34 among negative records. However, the negative group extends to 7438.80 nanomolar, producing greater dispersion and a larger separation between its mean and median. The median comparison is less influenced by these extreme values.

Within patient comparisons qualify the pooled pattern. Positive records have lower median affinity in four of six patients. Patients 12 and 16 show the opposite ordering. Thus, the answer to RQ1 is that positive records exhibit stronger predicted binding in the pooled sample, but the direction is not consistent across all patients.

The pattern is compatible with a role for binding characteristics in immune recognition, while the exceptions reinforce the distinction between binding and immunogenicity. Experimental work similarly distinguishes binding affinity from other properties of the peptide antigen complex (Harndahl et al., 2012).

5.3 Tumor Abundance and Immunogenicity:

The second research question concerns tumor abundance. Positive records have a median abundance of 40.44 transcripts per million, compared with 11.38 among negative records. The median difference is 29.05 transcripts per million. Positive records therefore have approximately 3.55 times the median abundance of negative records.

The corresponding means are 39.87 and 32.68 transcripts per million. The smaller mean difference reflects the upper tail of the negative group, which reaches 460.08 transcripts per million. High transcript abundance therefore occurs among candidates without validated immunogenicity and does not provide a sufficient classification rule.

Positive records have greater median abundance within four patients. Patients 10 and 16 show lower median abundance among positive records. The answer to RQ2 is consequently a positive pooled contrast with variation across patients, rather than a uniform relationship.

This distinction is consistent with evidence that expression contributes information about antigen presentation without directly establishing immune recognition (Abelin et al., 2017). The present comparison concerns transcript abundance and does not measure displayed peptide density.

5.4 Predicted Binding Stability and Immunogenicity:

The third research question concerns predicted complex persistence. Median stability is 3.28 hours among positive records and 2.10 among negative records. The difference is 1.18 hours, equivalent to 56.19 percent of the negative group median.

Mean stability is also greater among positive records, at 6.05 hours compared with 3.68 hours. Dispersion remains substantial: the positive group ranges from 0.14 to 49.34 hours, while the negative group ranges from 0.12 to 55.36 hours. Longer predicted persistence characterizes the positive group on average but does not uniquely identify its members.

Positive records have greater median stability in five of six patients. Patient 12 shows the opposite ordering. Stability therefore exhibits the most consistent within patient direction among the three features, although this count does not establish superior predictive performance.

The answer to RQ3 is that positive records have longer predicted stability in the pooled sample and in most patient comparisons. This pattern is compatible with the relevance of complex persistence identified in stability prediction research (Rasmussen et al., 2016).

5.5 Patient Variation and Sensitivity of the Descriptive Findings:

The within patient results reveal variation that pooled summaries conceal. Patient 12 departs from the pooled affinity and stability patterns, while patient 16 departs from the pooled affinity and abundance patterns.

Patient 10 departs from the pooled abundance pattern. These observations show that a common pooled direction can coexist with opposing patient specific contrasts.

Table 6: Within patient median comparisons

Patient	Affinity Positive	Affinity Negative	Abundance Positive	Abundance Negative	Stability Positive	Stability Negative
1	20.35	54.05	73.26	32.70	3.84	1.84
2	19.10	38.40	53.51	33.19	7.49	3.49
3	16.15	49.00	38.53	25.23	3.70	1.59
10	24.40	212.20	2.45	2.87	2.11	0.56
12	204.80	115.00	8.56	5.85	0.82	1.81
16	160.00	94.30	1.28	4.40	3.97	2.14

An exploratory sensitivity check assigned equal weight to each patient by averaging the six patient specific medians within each outcome group. The resulting affinity summaries are 74.13 nanomolar for positive records and 93.83 for negative records. Abundance summaries are 29.60 and 17.37 transcripts per million, respectively, while stability summaries are 3.65 and 1.91 hours. All three directions remain consistent with the pooled comparisons. However, the affinity contrast becomes smaller, showing sensitivity of its magnitude to patient composition.

A separate exploratory check counted each patient and peptide HLA combination once. This reduced the sample from 404 biological sample records to 401 distinct combinations. The three removed repetitions all had negative outcomes. Negative group medians changed from 74.30 to 73.50 nanomolar for affinity, from 11.38 to 11.64 transcripts per million for abundance, and from 2.10 to 2.11 hours for stability. Positive group medians remained unchanged. The pooled directions therefore do not depend on these repeated complexes.

Spearman correlations identify partial overlap among the presentation features. Predicted affinity and stability have a correlation of negative 0.482. Affinity and abundance have a correlation of negative 0.097, while abundance and stability have a correlation of 0.087. Stronger predicted binding therefore tends to accompany longer predicted persistence, whereas abundance shows weaker pairwise correspondence with either feature. These coefficients describe record level relationships without establishing independent contributions to immunogenicity.

Across the three research questions, positive records show stronger median predicted binding, greater median abundance, and longer median predicted stability. These directions persist under the two exploratory sample weighting checks, but individual patients do not uniformly reproduce them. The evidence characterizes molecular features associated descriptively with the recorded immune outcome. It does not establish clinical vaccine benefit, which requires evidence beyond antigen presentation and recognition (Schumacher & Schreiber, 2015).

6. Discussion:

A pooled molecular signature can conceal differences between patients that matter for personalized vaccine target selection. Table 5 shows stronger predicted binding, greater transcript abundance, and longer predicted stability among records with validated immunogenicity. Table 6 qualifies this pattern by revealing opposing contrasts in individual patients. The added contribution of this analysis lies in distinguishing pooled separation from patient level consistency, rather than claiming discovery of the presentation features themselves. This distinction complements frameworks that integrate antigen presentation and recognition when evaluating tumor immune susceptibility (Łuksza et al., 2017). The findings show that an aggregate pattern should not automatically become a uniform selection rule for every patient.

The observed contrasts are compatible with complementary biological constraints, but their overlap prevents a deterministic interpretation. Table 5 includes negative records with strong predicted binding, high abundance, or long stability, alongside positive records with less favorable individual characteristics. Presentation features therefore identify tendencies within the assessed candidates rather than sufficient conditions for recognition. Strønen et al. (2016) demonstrated that donor T cells could recognize candidate neoantigens neglected by patients' own T cell populations, illustrating why peptide characteristics and immune recognition need not coincide. The present analysis adds a descriptive boundary to feature based selection: favorable molecular values do not uniquely determine the recorded immune outcome.

Consistency across patients provides a different perspective from the magnitude of pooled separation. Predicted stability follows the pooled direction in five patients, whereas affinity and abundance do so in four. Table 6 therefore identifies stability as the most directionally consistent feature in this sample, without establishing that it dominates prediction or explains the largest biological contribution. The moderate inverse correlation between affinity values and stability also indicates shared information between binding strength and persistence. These findings refine the interpretation of presentation features as partly related constraints rather than interchangeable indicators. They support the integrated selection perspective described by Hu et al. (2018),

while showing why the largest pooled contrast and the most consistent patient level pattern need not identify the same feature.

Sample construction changes the empirical representation of immunogenicity. Removing records with unavailable abundance raises the proportion of positive outcomes and alters patient contributions, making completeness part of the interpretation rather than a neutral administrative step. Equal patient weighting preserves all three descriptive directions but reduces the affinity contrast, whereas counting repeated complexes once produces little change. Together with Tables 5 and 6, these checks distinguish sensitivity to patient composition from sensitivity to repeated biological samples. This matters for translating candidate assessment into vaccine design because clinical feasibility and immunogenicity depend on a sequence of selection and experimental decisions, as illustrated by Ott et al. (2017). Transparent reporting of those decisions helps separate candidate characteristics from the composition of the evaluable sample.

The international relevance lies in a transferable evaluation principle: pooled candidate differences should be examined alongside patient variation and sample selection before informing personalized target prioritization. Table 5 supports considering all three presentation features, while Table 6 argues against treating their pooled ordering as universal. Vaccine development reviews emphasize the interconnected challenges of candidate identification, immune response assessment, and individualized production (Blass & Ott, 2021). This analysis contributes a focused empirical perspective by showing that descriptive direction can persist under alternative weighting while its magnitude changes. For research teams comparing selection pipelines, the practical implication is to report patient specific contrasts and coverage losses alongside pooled summaries, preserving the distinction between molecular prioritization and demonstrated therapeutic benefit.

7. Conclusion and Implications:

Personalized cancer vaccine selection loses precision when pooled evidence obscures differences between patients. This analysis shows that favorable molecular profiles distinguish validated candidates in aggregate, while individual patients can display opposing patterns. The contribution is an empirical distinction between directional persistence across alternative descriptive summaries and consistency within patients. This distinction qualifies presentation based reasoning by showing that pooled separation need not imply a uniform patient specific ordering. The evidence therefore supports treating molecular characteristics as complementary selection information rather than sufficient criteria for immune recognition. For research managers, the implication is to assess candidate rankings alongside patient coverage before allocating experimental resources. Operational workflows should preserve biological sample identifiers, document missing measurements, and compare pooled summaries with patient specific patterns. Research governance should encourage transparent reporting of exclusions and repeated observations so that candidate selection remains traceable. These practices can improve comparability across research teams without imposing an unsupported universal threshold. Their international relevance rests on a transferable evaluation principle: evidence used for personalized decisions should retain the variation that personalization seeks to address. The practical value lies in making prioritization more accountable, while keeping conclusions about experimental recognition separate from claims about therapeutic benefit.

The dataset offers opportunities for broader validation through additional patients, cancer types, and independently assessed candidates. Future analyses should examine whether the observed patterns persist with more complete transcript measurements and balanced patient coverage. Independent experimental measurements could assess uncertainty introduced by molecular predictions. Prospective evaluations should compare selection strategies using patients excluded from model development and distinguish candidate recognition from clinical response. Repeated sampling could examine whether candidate characteristics and recognition change over treatment. These extensions would test transferability and clarify the biological processes underlying the descriptive patterns identified here.

References:

1. Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G. L., Eisenhaure, T. M., Clauser, K. R., Hacohen, N., Rooney, M. S., Carr, S. A., & Wu, C. J. 2017. Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. Immunity, 46, issue 2, 315-326. https://doi.org/10.1016/j.immuni.2017.02.007
2. Blass, E., & Ott, P. A. 2021. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. Nature Reviews Clinical Oncology, 18, issue 4, 215-229. https://doi.org/10.1038/s41571-020-00460-2
3. Harndahl, M., Rasmussen, M., Roder, G., Pedersen, I. D., Sørensen, M., Nielsen, M., & Buus, S. 2012. Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity. European Journal of Immunology, 42, issue 6, 1405-1416. https://doi.org/10.1002/eji.201141774
4. Hu, Z., Ott, P. A., & Wu, C. J. 2018. Towards personalized, tumour-specific, therapeutic vaccines for cancer. Nature Reviews Immunology, 18, issue 3, 168-182. https://doi.org/10.1038/nri.2017.131

5. Łuksza, M., Riaz, N., Makarov, V., Balachandran, V. P., Hellmann, M. D., Solovyov, A., Rizvi, N. A., Merghoub, T., Levine, A. J., Chan, T. A., Wolchok, J. D., & Greenbaum, B. D. 2017. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. Nature, 551, issue 7681, 517-520. https://doi.org/10.1038/nature24473
6. Ott, P. A., Hu, Z., Keskin, D. B., Shukla, S. A., Sun, J., Bozym, D. J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L., Chen, C., Olive, O., Carter, T. A., Li, S., Lieb, D. J., Eisenhaure, T., Gjini, E., Stevens, J., Lane, W. J., . . . Wu, C. J. 2017. An immunogenic personal neoantigen vaccine for patients with melanoma. Nature, 547, issue 7662, 217-221. https://doi.org/10.1038/nature22991
7. Rasmussen, M., Fenoy, E., Harndahl, M., Kristensen, A. B., Nielsen, I. K., Nielsen, M., & Buus, S. 2016. Pan-specific prediction of peptide-MHC class I complex stability, a correlate of T cell immunogenicity. The Journal of Immunology, 197, issue 4, 1517-1524. https://doi.org/10.4049/jimmunol.1600582
8. Schumacher, T. N., & Schreiber, R. D. 2015. Neoantigens in cancer immunotherapy. Science, 348, issue 6230, 69-74. https://doi.org/10.1126/science.aaa4971
9. Strønen, E., Toebes, M., Kelderman, S., van Buuren, M. M., Yang, W., van Rooij, N., Donia, M., Böschen, M. L., Lund-Johansen, F., Olweus, J., & Schumacher, T. N. 2016. Targeting of cancer neoantigens with donor-derived T cell receptor repertoires. Science, 352, issue 6291, 1337-1341. https://doi.org/10.1126/science.aaf2288
10. Wells, D. K., van Buuren, M. M., Dang, K. K., Hubbard-Lucey, V. M., Sheehan, K. C. F., Campbell, K. M., Lamb, A., Ward, J. P., Sidney, J., Blazquez, A. B., Rech, A. J., Zaretsky, J. M., Comin-Anduix, B., Ng, A. H. C., Chour, W., Yu, T. V., Rizvi, H., Chen, J. M., Manning, P., . . . Defranoux, N. A. 2020. Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. Cell, 183, issue 3, 818-834.e13. https://doi.org/10.1016/j.cell.2020.09.015
11. Yadav, M., Jhunjhunwala, S., Phung, Q. T., Lupardus, P., Tanguay, J., Bumbaca, S., Franci, C., Cheung, T. K., Fritsche, J., Weinschenk, T., Modrusan, Z., Mellman, I., Lill, J. R., &Delamarre, L. 2014. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. Nature, 515, issue 7528, 572-576. https://doi.org/10.1038/nature14001