

# Demographic distribution matching between real-world and virtual phantom population

Dhrubajyoti Ghosh<sup>1</sup> | Fakrul Tushar<sup>2</sup> | Lavsén Dahal<sup>2</sup> | Liesbeth Vancoillie<sup>2</sup> | Kyle J. Lafata<sup>2</sup> | Ehsan Samei<sup>2</sup> | Joseph Y. Lo<sup>2</sup> | Sheng Luo<sup>1</sup>

<sup>1</sup>Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, USA

<sup>2</sup>Center for Virtual Imaging Trials, Carl E. Ravin Advanced Imaging Laboratories, Department of Radiology, Duke University School of Medicine, Durham, North Carolina, USA

## Correspondence

Dhrubajyoti Ghosh, Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA.

Email: [dg302@duke.edu](mailto:dg302@duke.edu)

## Funding information

National Institutes of Health; NIBIB, Grant/Award Number: P41 EB028744

## Abstract

**Background:** The adoption of virtual imaging trials (VITs) is rapidly expanding, offering a cost-effective and ethically viable alternative to large-scale clinical trials for imaging system evaluation. However, differences in demographic composition between virtual phantom populations and real-world clinical cohorts can introduce bias in imaging performance assessments, particularly for underrepresented populations. Such discrepancies, if unaddressed, can limit the translational relevance of VIT findings by misrepresenting diagnostic performance across diverse patient groups.

**Purpose:** To address this limitation, we introduce DISTINCT (Distributional Sub-sampling for Covariate-Targeted Alignment), a statistical framework for selecting demographically aligned subsamples from large clinical datasets to support robust comparisons with virtual cohorts.

**Methods:** We applied DISTINCT to the National Lung Screening Trial (NLST) and a companion virtual trial dataset (VLST). The algorithm jointly aligned typical continuous (age, BMI) and categorical (sex, race, ethnicity) variables by constructing multidimensional bins based on discretized covariates. For a given target size, DISTINCT samples individuals to match the joint demographic distribution of the reference population. We evaluated the demographic similarity between VLST and progressively larger NLST subsamples using Wasserstein and Kolmogorov–Smirnov (K-S) distances to identify the maximal subsample size with acceptable alignment. After demographic alignment, we evaluated lung cancer risk prediction performance by applying two established NLST risk scores to the aligned subsamples and assessing their stability with receiver operating characteristic (ROC) analysis.

**Results:** The DISTINCT algorithm identified a maximal demographically aligned NLST subsample of 9974 participants that preserved similarity to the VLST population. To assess whether such aligned subsets were sufficient for downstream applications, we applied two established NLST lung cancer risk scores and evaluated their performance using ROC analysis. Area under the curve (AUC) estimates stabilized once subsample sizes exceeded approximately 6000 participants, demonstrating that moderately sized aligned subsets provide reliable predictive model evaluation. Stratified analyses revealed demographic-specific variations in AUC, underscoring the importance of covariate alignment for fair and representative comparisons.

**Conclusion:** DISTINCT provides a statistically rigorous and scalable approach for covariate alignment between real and virtual imaging cohorts based on demographic factors of variability. Although demonstrated for lung cancer screening with low-dose CT, the framework is broadly applicable to other imaging modalities and diseases, and across wide ranges of factors of

variability. By enabling fair and representative performance assessments, DISTINCT advances the integration of VITs into imaging research and protocol optimization workflows.

#### KEYWORDS

demographics matching, virtual clinical trials, Wasserstein distance

## 1 | INTRODUCTION

Lung cancer remains a leading cause of cancer mortality worldwide, underscoring the critical need for effective early detection and screening strategies.<sup>1–3</sup>

Low-dose computed tomography (LDCT) has significantly improved early detection, motivated by the National Lung Screening Trial (NLST), which demonstrated reduced mortality among high-risk individuals.<sup>4</sup> However, it has been suggested that the demographic homogeneity of the NLST cohort, predominantly white participants, limits the generalizability of its findings across more diverse populations.<sup>5,6</sup> Demographic biases in terms of anatomical variability associated with age, sex, and body mass index (BMI) can influence diagnostic performance. Failure to account for these differences can lead to biased image-based models, dose factors, and clinical recommendations, ultimately exacerbating disparities in underrepresented patient populations.

To complement traditional clinical trials, virtual imaging trials (VITs) have emerged as powerful tools. By using computationally generated human phantoms,<sup>7–9</sup> VITs enable efficient simulation of patient anatomy, disease progression, and imaging protocols.<sup>10–12</sup> These trials circumvent many logistical, financial, and ethical constraints associated with large-scale human studies. The Virtual Lung Screening Trial (VLST) exemplifies this approach, leveraging high-fidelity phantoms to systematically evaluate LDCT performance, protocol optimization, and image quality under controlled conditions.<sup>13–15</sup> The VLST dataset comprises virtual phantoms generated from de-identified chest CT scans using the XCAT framework, which models organ-level anatomy and tissue composition to preserve realistic intersubject variability while retaining associated demographic metadata such as age, sex, race, and ethnicity. In the current implementation, race and ethnicity metadata inform cohort composition but do not directly alter individual phantom anatomy. However, phantom populations can differ demographically from real-world cohorts such as the NLST, potentially introducing confounding factors. Demographic variables like age and BMI affect radiation dose deposition, lung nodule detectability, and reconstruction accuracy, which are core concerns in image fidelity. Therefore, aligning the demographic composition of virtual and clinical populations is essential for unbiased and generalizable conclusions.

Addressing these discrepancies requires statistical approaches capable of capturing full distributional differences rather than focusing solely on central tendencies. Conventional methods such as *t*-tests or Wilcoxon tests assess mean or median differences but overlook distributional variation. In contrast, more comprehensive metrics, such as the Kolmogorov–Smirnov (K-S) and Wasserstein distances, quantify differences between entire distributions. The K-S distance measures the maximum divergence between empirical cumulative distribution functions (ECDFs),<sup>16</sup> while the Wasserstein distance, rooted in optimal transport theory, quantifies the total “effort” required to morph one distribution into another.<sup>17,18</sup> These measures offer more detailed insights when assessing population-level comparability in imaging trials.

Currently, there is no standardized methodological framework for ensuring demographic comparability between virtual phantoms and clinical cohorts. While several methods exist for aligning two clinical cohorts, such as propensity score matching, inverse probability weighting, and entropy balancing, these primarily target mean balance or low-order moments and fail to align the full joint distribution of covariates. This gap limits the translational relevance of VIT findings. To address this challenge, we introduce DISTINCT (Distributional Subsampling for Covariate-Targeted Alignment), a statistical algorithm that selects demographically aligned subsamples from large clinical datasets to match other populations, including virtual populations. Traditional mean-based comparisons such as *t*-tests or Wilcoxon tests assess only central tendency and can miss substantial differences in spread or shape; equality of means therefore does not imply comparable populations. To capture full distributional discrepancies relevant to imaging variability across scanners and cohorts, DISTINCT employs K-S and Wasserstein distances, widely used in probability metrics and recently in imaging harmonization.<sup>16–20</sup> This formal distributional alignment extends earlier virtual-trial approaches that relied primarily on descriptive summaries.<sup>10</sup> This alignment improves the external validity of VIT results, supporting reliable assessments of imaging protocols, radiation dose strategies, and diagnostic performance across diverse populations. While we focus on LDCT for lung cancer screening, DISTINCT is broadly applicable to other variabilities, imaging modalities, clinical domains and certain factors

of variability, facilitating equitable and representative virtual trials.

In this study, we treat VLST as the target distribution and NLST as the source cohort from which DISTINCT extracts matched subsamples. DISTINCT is agnostic to the direction of matching and can be applied with either cohort serving as the target distribution. When large virtual populations are available or can be generated at scale, DISTINCT can instead be used to construct virtual samples whose joint covariate distribution matches a specified clinical target. Here, we adopt the NLST-to-VLST alignment because the publicly available VLST release is substantially smaller than the NLST cohort. In this finite virtual-cohort regime, explicit distributional matching is a critical methodological step: Without first ensuring comparability of covariate distributions, subsequent evaluations (e.g., comparisons of risk-model behavior or protocol-dependent endpoints) may reflect population mismatch rather than the scientific factors under study.

The remainder of this article is structured as follows: Section 2 details our methodology, including Section 2.1 on demographic alignment metrics, Section 2.2 on the DISTINCT algorithm, and Section 2.3 on demographic discrepancies between VLST and NLST. Section 3 presents results of applying DISTINCT to align NLST with VLST. Section 4 discusses implications for imaging physics and outlines future directions, and Section 5 concludes the article.

## 2 | METHODS

To ensure a rigorous and quantitative assessment of demographic differences between the NLST and the VLST populations, we employed advanced statistical techniques to measure distributional divergence and developed a novel algorithm to facilitate demographically aligned dataset comparisons. This section presents the statistical metrics used to assess variability, outlines the DISTINCT algorithm for generating matched subsamples, and describes the procedure for aligning virtual and clinical trial populations. Throughout, we treat VLST as the fixed target distribution and sample from NLST to match it. This choice is driven by the present sample sizes (NLST: 26,722; VLST: 264) and by the goal of identifying the largest clinically observed subset that is jointly comparable to the available virtual cohort. DISTINCT is not inherently directional; rather, directionality here reflects the practical regime in which the virtual cohort is currently smaller than the clinical cohort.

### 2.1 | Statistical metrics for assessing distributional differences

In medical physics research, accurately comparing demographic distributions requires statistical tools that

assess more than just central tendency. Common approaches such as the  $t$ -test or Wilcoxon rank-sum test compare means or medians, but may fail to detect broader differences in shape, skewness, or modality, features that directly impact image quality, radiation dose, and model generalizability. To capture these broader patterns, we use two complementary distributional distance metrics: the K-S distance and the Wasserstein distance.

The K-S distance is a nonparametric measure that quantifies the maximum absolute difference between the eECDFs of two samples. For two empirical distributions  $F_A(x)$  and  $F_B(x)$ , the K-S distance is defined as follows:

$$d_{KS}(A, B) = \sup_x |F_A(x) - F_B(x)|.$$

To assess statistical significance, the associated  $p$ -value is approximated by the following:

$$P(d_{KS}(A, B) > t) = Q_{KS} \left( \sqrt{\frac{n_A n_B}{n_A + n_B}} \cdot t \right),$$

where  $Q_{KS}$  is the complementary cumulative distribution function of the Kolmogorov distribution, and  $n_A, n_B$  denote the sample sizes. The K-S test is particularly sensitive to differences in both location and shape.

The Wasserstein distance, also known as the Earth mover's distance, is derived from optimal transport theory. It measures the minimal cost required to transform one probability distribution into another by shifting probability mass across the sample space. The 1-Wasserstein distance for univariate distributions is given by the following:

$$d_W(A, B) = \int_0^1 |F_A^{-1}(p) - F_B^{-1}(p)| dp,$$

where  $F^{-1}$  represents the quantile function. Unlike the K-S metric, the Wasserstein distance accounts for the magnitude of differences throughout the entire distribution, providing a more comprehensive summary of dissimilarity.

To evaluate the significance of the observed Wasserstein distance, we use a permutation-based method. This involves pooling the two datasets, randomly shuffling the group labels, and computing Wasserstein distances under the null hypothesis that the samples are from the same distribution. The empirical  $p$ -value is computed as follows:

$$P(d_W > t) = \frac{1 + \#\{d_{W_j}^\pi > t\}}{1 + m},$$

where  $d_{W_j}^\pi$  denotes the Wasserstein distance computed under the  $j$ th permutation and  $m$  is the number of permutations performed.

Both the K-S and Wasserstein distances are well-established measures of distributional similarity that have been widely applied across scientific domains. In machine learning, Wasserstein-based metrics underpin generative adversarial networks (GANs) and domain adaptation frameworks for aligning data distributions.<sup>21,22</sup> Similarly, K-S and related distance measures have been used in epidemiology and public health to compare exposure or biomarker distributions across demographic subgroups.<sup>23–25</sup> DISTINCT adapts these general-purpose metrics to the specific context of VITs, where quantifying alignment between virtual and clinical populations is essential for validation and generalizability.

These two statistical metrics provide robust and interpretable assessments of demographic divergence between the NLST and VLST populations. They form the foundation for the matching strategy described in the following section.

## 2.2 | Demographic matching via the DISTINCT algorithm

To reduce confounding from demographic imbalance between study populations, we introduce the DISTINCT (Distribution Integrator for Statistical Consistency and Normativity Technique) algorithm. DISTINCT constructs a subsample from a larger dataset that matches the demographic composition of a smaller target population. It accommodates both continuous and discrete demographic variables by discretizing continuous features and combining all variables into a unified binning structure for stratified sampling.

Let  $v_C$  represent the values for a set  $C$  of continuous variables (e.g., age, BMI), and let  $v_D$  denote values for a set  $D$  of discrete variables (e.g., sex, race, ethnicity). Each continuous variable  $v_C \in v_C$  is discretized into  $g$  bins using a binning function  $f_g(v_C)$ , yielding integer-valued labels. The resulting discretized vector is  $f_g(v_C) = \{f_g(v_C) \mid c \in C\}$ . We then concatenate  $v_D$  and  $f_g(v_C)$  to form a unified demographic label for each individual.

As an example, a subject who is female (label: 0), Non-Hispanic (label: 1), Asian (label: 3), with age 62 and BMI 27, would receive the demographic label (0, 1, 3, 2, 3) if age and BMI are binned into intervals (55–60, 60–65, 65–70, 70–75) and (10–18.5, 18.5–25, 25–30, 30+), respectively (Figure 1).

The DISTINCT algorithm proceeds as follows:

1. **Define demographic bins.** Combine discretized continuous and categorical variables into unified bin labels. Each unique label defines a demographic stratum.
2. **Compute target bin proportions.** For each bin  $l$ , compute  $p_l = y_l/N_T$ , where  $y_l$  is the number of individuals

in bin  $l$  in the target dataset, and  $N_T$  is the total size of the target dataset.

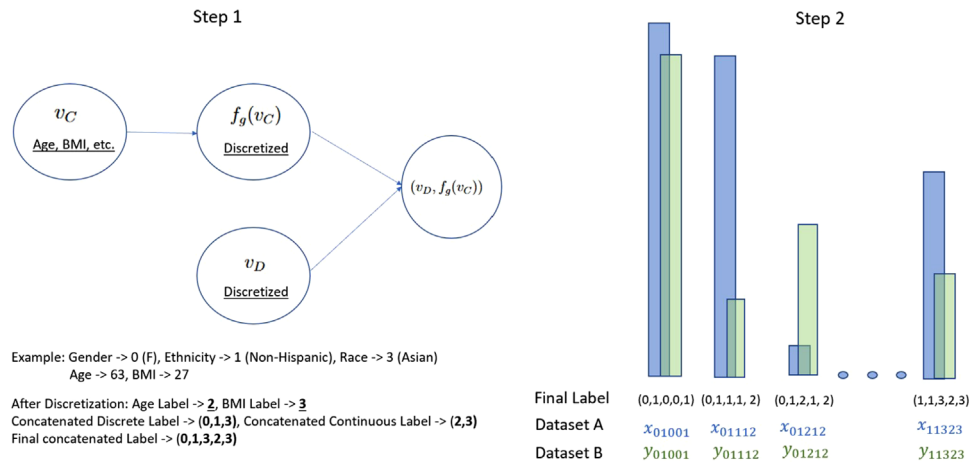
3. **Sample from the larger population.** For a given proposed subsample size  $N$ , calculate the target count per bin as  $\lfloor Np_l \rfloor$ . If the larger dataset contains at least that many individuals in bin  $l$  (i.e.,  $x_l \geq \lfloor Np_l \rfloor$ ), sample  $\lfloor Np_l \rfloor$  individuals at random; otherwise, include all  $x_l$  individuals.
4. **Construct the subsample.** Aggregate sampled individuals across all bins to form the final subsample. Note that due to bin constraints, the realized sample size may differ slightly from  $N$ .
5. **Assess demographic alignment.** Use the Wasserstein or K-S test to assess whether the subsample is statistically comparable to the target population in terms of demographic distribution.
6. **Iterate to maximize alignment.** If alignment is achieved, increase  $N$  and repeat the procedure to identify the largest possible aligned subset. If alignment fails, reduce  $N$  and repeat until a maximally sized demographically matched subset is identified.

As a concrete example, suppose the demographic feature space includes three discrete variables (two binary and one with five levels), and two continuous variables, each binned into five intervals (labeled 0–4). This yields  $2^2 \times 5^3 = 500$  distinct demographic bins. Let  $x_l$  denote the number of individuals in bin  $l$  in the larger dataset, and  $y_l$  the corresponding count in the target dataset. With target dataset size  $N_T$ , we compute  $p_l = y_l/N_T$ . For any proposed subsample size  $N$ , we aim to include  $\lfloor Np_l \rfloor$  individuals per bin. If  $x_l \geq \lfloor Np_l \rfloor$ , we sample  $\lfloor Np_l \rfloor$  individuals randomly; otherwise, we include all  $x_l$ .

This bin-wise sampling approach generates a subsample that closely mirrors the demographic distribution of the target dataset. Statistical tests verify alignment, and the algorithm iteratively adjusts  $N$  to identify the largest demographically compatible subset.

## 2.3 | Demographic comparison and subsampling for virtual clinical trials

Tushar et al.<sup>14</sup> introduced the VLST, which employs computationally generated virtual human phantoms derived from clinical CT/PET scans collected at Duke University Medical Center, a multihospital academic health system. The VLST dataset includes 264 virtual patients with a mean age of 59.53 years, 55.68% male, and 98.5% identifying as non-Hispanic. The average BMI is 27.13. In comparison, the NLST cohort used in this study consists of 26 722 participants drawn from the publicly available screening dataset. The cohort includes a broad age range (55–74 years), with approximately 59% male and 41% female participants, and about 97.4% identifying as non-Hispanic. Baseline demographic characteristics are



**FIGURE 1** Schematic overview of the DISTINCT algorithm. Demographic variables, both continuous (e.g., age, BMI) and discrete (e.g., sex, race, ethnicity), are discretized and combined to define stratification bins. The algorithm constructs a subsample from the larger dataset to match the demographic bin distribution of a smaller target dataset.

**TABLE 1** Baseline demographic characteristics of the NLST and VLST cohorts.

	NLST (LDCT, $N = 26\,722$ )	VLST ( $N = 264$ )
Age (years)	61.42 ± 5.03 (43–75)	59.53 ± 14.44 (2–86)
Female—no. (%)	10 953 (40.98)	117 (44.32)
Race—no. (%)	White: 24 289 (90.9) Black: 1195 (4.5) Asian: 559 (2.1) Other/Unknown: 687 (2.5)	White: 198 (75.0) Black: 56 (21.2) Asian: – Other/Unknown: 10 (3.8)
Non-Hispanic—no. (%)	26 079 (97.4)	260 (98.5)

summarized in Table 1, while distributions of age and BMI in both the VLST and NLST cohorts are shown in Figure 2.

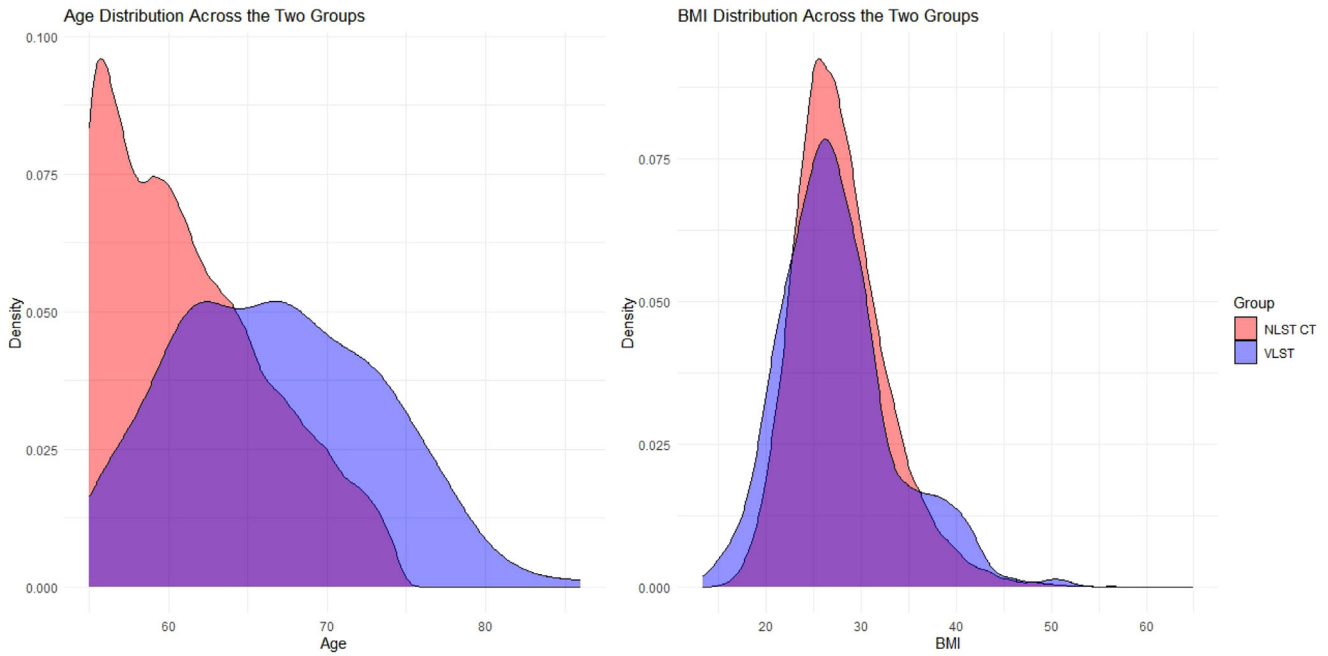
The VLST dataset<sup>14</sup> comprised 264 virtual patients generated from de-identified chest CT scans acquired at the Duke University Health System using the XCAT framework. Each virtual subject was modeled from a real CT case to produce anatomically realistic phantoms, 154 of which included simulated nodules. Demographic attributes such as age, sex, BMI, race, and ethnicity were inherited from the corresponding source patient-level metadata when available. Race and ethnicity labels were used only to define population-level composition and did not alter individual phantom anatomy. The phantom-generation pipeline and accompanying cohort-level metadata (age, sex, BMI, race, ethnicity) are documented and provided with the dataset.<sup>14</sup> Because VLST and NLST originate from independent cohorts, our analysis focuses on population-level similarity rather than one-to-one correspondence. The Hispanic subgroup was small in both VLST (about 1.5%) and NLST (about 2.6%), so results for this group are descriptive only.

Statistical comparisons between the NLST and VLST populations using the Wasserstein test reveal marked distributional differences. Specifically, the  $p$ -values for

age, race, and BMI are  $< 10^{-4}$ ,  $< 10^{-3}$ , and  $< 10^{-3}$ , respectively, while ethnicity and sex show less pronounced discrepancies ( $p = 0.103$  and  $0.283$ , respectively). These results highlight substantial demographic mismatches that may confound downstream imaging and modeling analyses if not properly addressed.

To enable meaningful statistical comparisons and eliminate confounding, we use the DISTINCT algorithm to extract demographically aligned subsamples from the NLST dataset that match the VLST distribution. Given that the real-world NLST cohort ( $N = 26\,722$ ) is much larger than the VLST cohort ( $N = 264$ ), subsampling from NLST is computationally feasible and appropriate in the current setting. While it would be ideal to sample from an expanded virtual cohort, current computational limits constrain the size of VLST. However, with ongoing phantom development, the VLST is expected to eventually exceed the size of most clinical datasets, at which point demographic alignment in the reverse direction will be feasible.

The DISTINCT framework is designed to operate flexibly regardless of which population is larger, making it suitable for both current and future demographic alignment tasks. In this analysis, we apply DISTINCT to extract NLST subsamples that closely mirror the VLST demographic profile, thereby enabling statistically valid



**FIGURE 2** Histograms of age (lower truncated at 55) and BMI for NLST CT and VLST cohorts, showing significant distributional differences between the two populations.

and demographically balanced comparisons of imaging protocols and modeling outcomes.

As a secondary analysis, we evaluated how demographic alignment affects the behavior of established lung cancer risk scores within the NLST cohort. Specifically, we considered two scores originally developed for the NLST Spiral CT arm by Pinsky et al.<sup>26</sup>: a radiologist recommendation-based propensity score (PSFR) and a nodule-size-based score (PSSZ). PSFR summarizes radiologists' follow-up recommendations and associated imaging findings into a single malignancy risk index, whereas PSSZ is defined solely from the maximum diameter of the largest noncalcified nodule (NCN) recorded at baseline CT. In this study, PSFR and PSSZ values, along with lung cancer outcomes, were taken directly from the NLST dataset. DISTINCT-aligned NLST subsamples were then used to examine how the AUC of these scores varies with sample size and demographic composition.

### 3 | RESULTS

We applied the DISTINCT algorithm to generate demographically aligned subsamples from the NLST dataset that closely match the demographic structure of the VLST population. Given the substantial difference in sample sizes between the two cohorts (NLST: 26 722 participants in the LDCT arm; VLST: 264 virtual phantoms), we extracted a series of progressively larger subsamples from the NLST cohort. At each subsample

size, we evaluated demographic similarity using formal statistical tests. This iterative procedure enabled us to determine the largest NLST subsample that exhibited no statistically significant differences from the VLST population across key demographic variables.

The demographic alignment process accounted for continuous variables (age and BMI) and categorical variables (sex, race, and ethnicity). To ensure computational tractability and consistent binning across cohorts, age was grouped into 5-year intervals starting from 55 years, while BMI was categorized into standard clinical ranges: underweight (<18.5), normal weight (18.5–24.9), overweight (25.0–29.9), and obese ( $\geq 30$ ). Categorical variables retained their native levels, including binary (e.g., sex) and multilevel (e.g., race) categories.

The DISTINCT algorithm defined multidimensional demographic bins by combining discretized continuous variables (age and BMI) with categorical variables (sex, race, ethnicity). For each bin  $\ell$ , the proportion of individuals in VLST was used to define the target sampling proportion,  $p_\ell = y_\ell / N_T$ , where  $y_\ell$  denotes the number of individuals in bin  $\ell$  and  $N_T = 264$  is the total VLST cohort size. Given a proposed subsample size  $N$ , the algorithm selected approximately  $\lfloor Np_\ell \rfloor$  individuals from each bin. If the number of available individuals  $x_\ell$  in NLST exceeded the target, a random sample of size  $\lfloor Np_\ell \rfloor$  was drawn; otherwise, all  $x_\ell$  individuals were included. This bin-wise procedure was repeated across increasing values of  $N$  to identify the maximal NLST subsample size for which demographic similarity could be maintained.

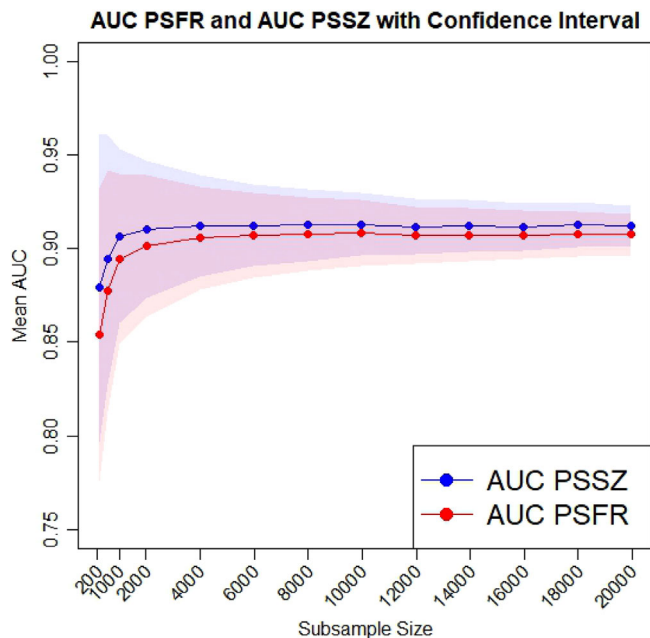
**TABLE 2**  $p$ -values for Wasserstein and K-S tests comparing demographic distributions between VLST (fixed at  $N = 264$ ) and progressively larger subsamples drawn from NLST. Alignment is considered adequate if all  $p$ -values exceed 0.05.

Test	Subsample size	Age	Sex	Race	Ethnicity	BMI
Wasserstein	279	0.983	0.757	0.725	0.676	0.904
	559	0.610	0.968	0.854	0.834	0.871
	1038	0.344	0.563	0.895	0.413	0.388
	2019	0.960	0.856	0.750	0.663	0.437
	3998	0.688	0.744	0.990	0.977	0.786
	5981	0.837	0.747	0.783	0.644	0.772
	7981	0.259	0.909	0.919	0.738	0.193
	9974	0.533	0.604	0.368	0.303	0.256
	11 963	0.579	0.494	0.442	0.210	< 1e-3
	13 963	0.213	0.452	0.058	0.114	< 1e-3
	15 965	0.074	0.176	< 1e-3	0.102	< 1e-3
	17 958	0.021	0.143	< 1e-3	0.084	< 1e-3
	K-S	279	0.933	0.951	0.952	1.000
559		0.993	0.746	0.979	0.442	0.296
1038		0.877	0.308	0.831	0.312	0.821
2019		0.775	0.924	0.614	0.205	0.380
3998		0.664	0.642	0.619	0.518	0.134
5981		0.749	0.714	0.757	0.619	0.101
7981		0.538	0.514	0.968	0.152	0.062
9974		0.233	0.706	0.053	0.314	0.052
11 963		0.310	0.124	< 1e-3	0.218	0.011
13 963		0.058	0.258	< 1e-3	0.198	< 1e-3
15 965		< 1e-3	0.312	< 1e-3	0.252	< 1e-3

To assess demographic similarity, we applied both the Wasserstein distance and the K-S test to each variable independently. Table 2 reports the resulting  $p$ -values for each metric at various subsample sizes. Alignment was considered acceptable if all  $p$ -values exceeded 0.05, indicating no significant distributional differences. This criterion was satisfied up to a subsample size of approximately 9974, beyond which age and BMI began to diverge significantly. At smaller sizes (e.g., 279 and 559), the algorithm achieved excellent alignment across all demographic variables, as reflected by nonsignificant  $p$ -values. These findings are expected: Smaller subsamples offer greater flexibility in matching the target distribution, even in the presence of baseline population-level differences. The maximal aligned size is determined empirically by solving the subsampling problem under the observed joint distribution of age, sex, BMI, race, and ethnicity. In general, no closed-form expression exists for this limit because it depends on finite cell counts across all multidimensional covariate combinations and the integer constraints imposed by sampling. Consequently, the attainable size must be evaluated algorithmically rather than derived analytically.

As subsample size increased, maintaining distributional similarity became more difficult due to inherent differences between the NLST and VLST populations. Sex alignment was maintained across all sample sizes, reflecting the similar female proportions in VLST (44.3%) and NLST (41%). In contrast, BMI emerged as the most limiting factor. For example, in the Wasserstein analysis, BMI alignment was rejected at  $N = 11\,963$  ( $p < 0.001$ ). Race and ethnicity also showed decreasing alignment, though less sharply. The K-S test exhibited similar trends, with BMI again being the first variable to exceed the significance threshold. Taking all variables into account, the largest subsample that maintained demographic alignment across all five dimensions was approximately 9974 individuals.

DISTINCT performs multivariate alignment by minimizing joint distributional distances across all demographic variables rather than optimizing each one independently. Thus, no single variable (e.g., BMI) is held fixed while the others vary; instead, the algorithm identifies a subset of participants whose combined age, sex, BMI, race, and ethnicity distributions most closely match those of the VLST population. The univariate plots illustrate marginal alignment behavior as



**FIGURE 3** AUC trajectories for PSFR and PSSZ across subsamples of increasing size. Points denote mean AUCs, and shaded bands indicate 95% confidence intervals. AUC values stabilize after approximately 6000 subjects.

sample size increases but do not represent separate optimizations. For instance, perfect BMI alignment could be achieved by fixing that dimension, but this would degrade alignment across the remaining variables because the optimization criterion is joint. The reported value of  $N = 9974$  corresponds to the largest NLST subsample for which all five variables were simultaneously aligned within the same participant set, based on joint Wasserstein and K-S statistics.

Following subsample generation, we assessed the impact of demographic matching and subsample size on predictive model performance. We focused on two established risk scores from the NLST study developed by Pinsky et al.<sup>26</sup>: the radiologist recommendation-based score (PSFR) and the nodule -size-based score (PSSZ). PSFR captures clinical judgment based on radiologists' follow-up recommendations, while PSSZ is an objective score derived solely from the maximum diameter of the largest NCN. For each NLST subsample, we computed the area under the receiver operating characteristic curve (AUC) for both scores using lung cancer status as the binary outcome. Figure 3 displays the AUC trajectories across subsample sizes. AUC estimates were highly variable at small sizes due to sampling noise, but they stabilized beyond approximately 6000 individuals. PSFR plateaued around 0.91, while PSSZ stabilized near 0.92. These results suggest that demographically aligned subsamples of moderate size suffice for robust model evaluation and that gains from larger cohorts may be marginal.

The stabilization of AUC with increasing aligned cohort size can also serve as a benchmark for evaluating the virtual population. Once demographic alignment is achieved, the AUC distribution from the matched NLST subset defines the expected performance envelope that a corresponding VLST analysis should reproduce, as a validation check of whether the virtual phantoms adequately represent the intended clinical population under demographic alignment. DISTINCT thereby links demographic alignment with performance validation, providing a quantitative bridge between population representativeness and diagnostic outcome fidelity.

To further examine subgroup-level model performance, we analyzed discrimination across demographic strata using the full NLST dataset. As detailed in Table A2 in the Appendix, females consistently showed higher AUCs than males for both PSFR (0.922 vs. 0.896) and PSSZ (0.934 vs. 0.895). Non-Hispanic participants also outperformed Hispanic individuals, with statistically significant differences in both scores. Lower BMI was associated with higher discrimination. These subgroup trends reinforce the importance of demographic alignment: Without appropriate matching, apparent differences in model performance may stem from differences in population composition rather than true disparities in model quality.

In summary, the DISTINCT algorithm enables the extraction of demographically aligned subsets from large real-world clinical trials, facilitating valid comparisons with virtual trial populations. The stability of model performance across matched subsets underscores the algorithm's utility for cross-cohort analysis and highlights the potential of virtual cohorts in translational cancer research.

## 4 | DISCUSSION

This study presents DISTINCT, a data-driven algorithm for subsampling a source population to closely match the joint distribution of covariates in a target population. DISTINCT offers a model-agnostic and nonparametric framework for demographic alignment. Unlike regression-based methods that rely on strong parametric assumptions or require labeled outcomes in both datasets, DISTINCT uses only covariate information from the target cohort to guide sample selection in the source cohort. It leverages multidimensional histogram-based binning, adaptive importance reweighting, and iterative subset construction to identify the largest feasible sample whose covariate distribution approximates that of the target population.

Empirical results using real-world clinical data from lung cancer screening demonstrate that DISTINCT generates demographically aligned subsamples that preserve downstream performance metrics, such as the area under the ROC curve (AUC) for established

risk prediction models. Unlike standard matching methods, which often focus on matching mean or median values or rely solely on pairwise matching, DISTINCT aligns the full joint distribution of continuous and categorical covariates using a combination of binning and distributional distance metrics. This approach ensures better global distributional alignment and avoids common issues of residual imbalance in higher moments or joint structures. Furthermore, the iterative design of DISTINCT identifies the largest subsample size that maintains statistical alignment with the target population, offering a more interpretable and scalable alternative to conventional matching or reweighting techniques, particularly when exact matching is infeasible in high-dimensional settings.

The motivation for this approach arose from the need to enable robust comparison of virtual and real-world imaging data, where discrepancies in population composition can confound estimates of diagnostic performance. In VITs, source data are often synthetically generated and may not reflect the diversity of real patient populations. This mismatch poses challenges for generalizability and fairness. By enabling demographic alignment, DISTINCT allows researchers to conduct trials under comparable population structures. More broadly, the method can be applied in a wide range of settings, including algorithm fairness auditing, synthetic data validation, external control arm construction, and domain adaptation in transfer learning.

In the present work, alignment was formulated in one direction, from the clinical population (NLST) toward the virtual population (VLST), because the virtual cohort is small and represents a specific subset of lung-screening participants. In broader applications of VITs, the desired direction of alignment is often reversed: The virtual population is generated or weighted to reflect the real-world clinical cohort that defines the intended use population. DISTINCT can readily accommodate this inverse formulation by optimizing the virtual cohort distribution to minimize its distance from the clinical reference. The current analysis therefore illustrates one use case of the framework, while the reverse alignment paradigm will become increasingly relevant as larger and more diverse virtual cohorts become available.

Our current formulation of DISTINCT assumes that all variables used for binning and subsampling are observed in both the source and target populations. In practice, missingness in covariates or differences in measurement platforms may necessitate imputation or dimension reduction strategies. While DISTINCT can be applied using a reduced subset of harmonized covariates, this may affect its ability to align more complex joint distributions including variability factors beyond demographics. Additionally, the choice of bin granularity and overlap thresholds affects both performance and sample size, suggesting a need for adaptive or data-driven tuning procedures. Future work may explore exten-

sions of DISTINCT for dynamic datasets, continuous recruitment, or longitudinal matching.

A further limitation is that our analysis was restricted to demographic variables (age, sex, BMI, race, ethnicity), which serve as indirect surrogates for anatomical and pathological features that ultimately determine CT image appearance. While these attributes are widely available, interpretable, and known to influence imaging outcomes, they do not fully capture morphometric variability such as lung volume, thoracic cavity dimensions, or nodule distribution. Quantitative image-derived features could provide a more direct and less biased basis for alignment between real and virtual cohorts. At present, however, such imaging biomarkers are not standardized for consistent extraction across both NLST and VLST datasets. The DISTINCT framework is inherently flexible and can incorporate such imaging-based covariates as standardized measures become available, providing an important direction for future development.

A related consideration is that DISTINCT was developed for empirically derived virtual phantoms, such as CT-based reconstructions that retain subject-specific anatomical variability. Its direct applicability to mathematically generated or procedurally defined phantoms, such as those representing anatomy using analytical primitives rather than measured data, may therefore be limited. Nonetheless, the framework could be extended if the generative parameters of such mathematical models were treated as random variables, enabling analogous distributional comparisons.

Another possible strategy for constructing virtual populations is to define demographic and physiological parameter distributions from a reference dataset and then generate virtual phantoms that conform to those specifications. This parameter-driven approach offers efficiency and control over the population composition, but it assumes that the selected parameters fully describe the complex dependencies between demographic factors and anatomical structure. If these relationships are incomplete or misspecified, the resulting cohort may be statistically representative yet anatomically unrealistic. DISTINCT provides a complementary, data-driven perspective by aligning existing real and virtual datasets based on observed distributions, preserving natural anatomical variability while achieving demographic comparability. Beyond validation, DISTINCT could also serve as a guiding tool for parameter-based generation: Its distributional distance metrics can identify underrepresented regions of the target population space and iteratively adjust the sampling of input parameters to minimize divergence from real-world distributions. In this way, DISTINCT can inform both the design and evaluation of virtual cohorts, offering a quantitative framework that balances simulation flexibility with empirical fidelity.

Together, these contributions position DISTINCT as a flexible tool for aligning populations across disparate

datasets in a way that preserves statistical comparability and enhances downstream inference. Its ability to operate without requiring outcome labels in the target population makes it particularly valuable in real-world scenarios where only covariate-level data are available. By providing a principled approach to demographic alignment, DISTINCT strengthens the foundation for fair and valid comparisons in translational and clinical research.

## 5 | CONCLUSION

DISTINCT provides a quantitative framework for aligning virtual and clinical populations by matching their joint demographic distributions using distributional distance metrics. Applied to the NLST and VLST cohorts, the method identified the largest subset of real patients that achieves simultaneous alignment across five key demographic variables, supporting fair and interpretable comparisons between virtual and clinical cohorts.

## ACKNOWLEDGMENTS

The study was supported by a grant from the National Institutes of Health NIBIB P41 EB028744.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## REFERENCES

1. Ferlay J, Colombet M, Soerjomataram I, et al. Cancer statistics for the year 2020: An overview. *Int J Cancer*. 2021;149(4):778-789.
2. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA-Cancer J Clin*. 2021;71(3):209-249.
3. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin*. 2023;73(1):17-48.
4. NLSTR Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;365(5):395-409.
5. Tammemägi MC, Katki HA, Hocking WG, et al. Selection criteria for lung-cancer screening. *N Engl J Med*. 2013;368(8):728-736.
6. NLSTR Team. Lung cancer incidence and mortality with extended follow-up in the National Lung Screening Trial. *J Thorac Oncol*. 2019;14(10):1732-1742.
7. Kainz W, Neufeld E, Bolch WE, et al. Advances in computational human phantoms and their applications in biomedical engineering—a topical review. *IEEE Trans Radiat Plasma Med Sci*. 2018;3(1):1-23.
8. Abadi E, Segars WP, Sturgeon GM, Harrawood B, Kapadia A, Samei E. Modeling “textured” bones in virtual human phantoms. *IEEE Trans Radiat Plasma Med Sci*. 2018;3(1):47-53.
9. Dahal L, Ghoghjhejad M, Vancoillie L, et al. XCAT 3.0: a comprehensive library of personalized digital twins derived from CT scans. *Med Image Anal*. 2025;103:103636.
10. Badano A, Graff CG, Badal A, et al. Evaluation of digital breast tomosynthesis as replacement of full-field digital mammography using an in silico imaging trial. *JAMA Netw Open*. 2018;1(7):e185474.
11. Abadi E, Segars WP, Tsui BM, et al. Virtual clinical trials in medical imaging: a review. *J Med Imaging*. 2020;7(4):042805.

12. Tushar FI, Dahal L, Sotoudeh-Paima S, et al. Data diversity and virtual imaging in AI-based diagnosis: a case study based on COVID-19. *arXiv preprint arXiv:2308.09730*. 2023.
13. Tushar FI, Vancoillie L, McCabe C, et al. Virtual NLST: towards replicating national lung screening trial. In: *Medical Imaging 2024: Physics of Medical Imaging*. Vol 12925. SPIE; 2024:442-447.
14. Tushar FI, Vancoillie L, McCabe C, et al. Virtual Lung Screening Trial (VLST): an in silico study inspired by the National Lung Screening Trial for lung cancer detection. *Med Image Anal*. 2025;103:103576.
15. Tushar FI, Vancoillie L, McCabe C, et al. Virtual imaging trials improved the transparency and reliability of AI systems in COVID-19 imaging. *arXiv e-prints*. arXiv:2308.09730. 2023.
16. Rachev ST. *Probability Metrics and the Stability of Stochastic Models*. Wiley; 1991.
17. Ramdas A, García Trillos N, Cuturi M. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*. 2017;19(2):47.
18. Villani C. *Optimal Transport: Old and New*. Vol 338. Springer; 2009.
19. Orhac F, Eertink JJ, Cottreau AS, et al. A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *J Nucl Med*. 2022;63(2):172-179.
20. Oh JH, Pouryahya M, Iyer A, Apte AP, Deasy JO, Tannenbaum A. A novel kernel Wasserstein distance on Gaussian measures: an application of identifying dental artifacts in head and neck computed tomography. *Comput Biol Med*. 2020;120:103731.
21. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: *Proceedings of the International Conference on Machine Learning*. PMLR; 2017:214-223.
22. Courty N, Flamary R, Tuia D, Rakotomamonjy A. Optimal transport for domain adaptation. *IEEE Trans Pattern Anal Mach Intell*. 2016;39(9):1853-1865.
23. Challen R, Dyson L, Overton CE, et al. Early epidemiological signatures of novel SARS-CoV-2 variants: establishment of B.1.617.2 in England. *MedRxiv*. 2021. doi:10.1101/2021.06.05.21258365
24. Adegboye OA, Kotze D. Epidemiological analysis of spatially misaligned data: a case of highly pathogenic avian influenza virus outbreak in Nigeria. *Epidemiol Infect*. 2014;142(5):940-949.
25. Wang S, Cai TT, Li H. Optimal estimation of Wasserstein distance on a tree with an application to microbiome studies. *J Am Stat Assoc*. 2021;116(535):1237-1253.
26. Pinsky PF, Gierada DS, Nath H, Kazerooni EA, Amorosa J. ROC curves for low-dose CT in the National Lung Screening Trial. *J Med Screen*. 2013;20(3):165-168.

**How to cite this article:** Ghosh D, Tushar F, Dahal L, et al. Demographic distribution matching between real-world and virtual phantom population. *Med Phys*. 2026;53:e70364. <https://doi.org/10.1002/mp.70364>

## APPENDIX A

### A.1 | Baseline comparability of LDCT and CXR arms in NLST

The National Lung Screening Trial (NLST) was a landmark clinical study designed to evaluate the effectiveness of low-dose helical computed tomography (LDCT) compared to standard chest radiography (CXR) in reducing lung cancer mortality. The trial enrolled over 53 000 high-risk individuals, primarily heavy smokers,

**TABLE A1** Baseline demographic characteristics of participants in the LDCT and CXR arms of the NLST. Statistical tests indicated no significant differences between the two groups across all variables assessed.

Characteristic	LDCT arm (N = 26 722)	CXR arm (N = 26 730)
Age (years)		
Mean $\pm$ SD (min–max)	61.42 $\pm$ 5.03 (43–75)	61.42 $\pm$ 5.02 (49–79)
Sex		
Female, no. (%)	10 953 (40.98%)	10 969 (41.05%)
Race		
White, no. (%)	24 289 (90.9%)	24 260 (90.8%)
Black/African American, no. (%)	1195 (4.5%)	1181 (4.4%)
Asian, no. (%)	559 (2.1%)	536 (2.0%)
Other/Unknown, no. (%)	687 (2.6%)	745 (2.8%)
Ethnicity		
Non-Hispanic, no. (%)	26 079 (97.4%)	26 039 (97.6%)

who were randomized to receive three annual screenings with either LDCT or CXR. The results of the NLST have had a lasting impact on lung cancer screening guidelines and early detection strategies.

Table A1 presents the baseline demographic characteristics of participants in the LDCT and CXR arms. To evaluate the comparability of the two groups, we assessed the distribution of key demographic variables, including age, sex, race, ethnicity, and body mass index (BMI). Statistical comparisons using the Wasserstein and Kolmogorov–Smirnov (K-S) tests revealed no significant differences between the two cohorts, with all  $p$ -values exceeding 0.05. This demographic equivalence supports valid cross-arm comparisons of outcomes such as diagnostic accuracy and ROC curves by reducing the likelihood of confounding from population structure.

## A.2 | Stratified AUC analysis of NLST risk scores by demographic subgroup

The NLST demonstrated that low-dose computed tomography (LDCT) screening reduced lung cancer mortality by 20% compared to CXR among individuals aged 55–74 years.<sup>4</sup> Although approximately 60% of NLST participants were male, the trial demonstrated comparable mortality reductions for both sexes.<sup>6</sup> The study population was predominantly White, raising concerns about generalizability to underrepresented racial and ethnic groups.<sup>5</sup>

To examine how predictive performance varies across demographic subgroups, we analyzed two lung cancer

**TABLE A2** Stratified area under the ROC curve (AUC) for PSFR and PSSZ risk scores across demographic subgroups in the NLST Spiral CT arm.

Demographics (N = 26722)	AUC		
	PSFR	PSSZ	
Age	55–60 (11 440)	0.909 $\pm$ 0.005	0.912 $\pm$ 0.005
	60–65 (8170)	0.903 $\pm$ 0.005	0.908 $\pm$ 0.005
	65–70 (4756)	0.901 $\pm$ 0.007	0.902 $\pm$ 0.007
	70–75 (2353)	0.911 $\pm$ 0.009	0.914 $\pm$ 0.009
Sex	Male (15 769)	0.896 $\pm$ 0.004	0.895 $\pm$ 0.005
	Female (10 953)	0.922 $\pm$ 0.004	0.934 $\pm$ 0.004
Ethnicity	Non-Hispanic (25 788)	0.931 $\pm$ 0.019	0.965 $\pm$ 0.012
	Hispanic (445)	0.905 $\pm$ 0.003	0.909 $\pm$ 0.003
BMI	10–18.5 (227)	0.901 $\pm$ 0.016	0.894 $\pm$ 0.012
	18.5–25 (7434)	0.920 $\pm$ 0.005	0.926 $\pm$ 0.005
	25–30 (11 143)	0.898 $\pm$ 0.005	0.900 $\pm$ 0.005
	>30 (7434)	0.892 $\pm$ 0.007	0.900 $\pm$ 0.007
Full dataset	0.910 $\pm$ 0.005	0.920 $\pm$ 0.005	

risk scores introduced by Pinsky et al.<sup>26</sup>: the radiologist-recommendation-based propensity score (PSFR) and the size-based propensity score (PSSZ). PSFR reflects radiologists' follow-up recommendations and incorporates subjective imaging features such as shape and density. In contrast, PSSZ is computed solely from the diameter of the largest noncalcified nodule (NCN), providing an objective and standardized risk measure.

Pinsky et al. reported that PSFR achieved a slightly higher AUC (AUC = 0.934) than PSSZ (AUC = 0.928), suggesting that radiologist judgment added incremental diagnostic value. Adjusting score thresholds improved specificity with limited sensitivity trade-offs. For example, a PSFR threshold of 3+ yielded 92.4% specificity and 86.9% sensitivity, while a PSSZ threshold of 8 mm resulted in 92.0% specificity and 83.2% sensitivity.

Table A2 reports AUC values stratified by demographic subgroup using the NLST Spiral CT data. Females exhibited higher AUCs than males for both PSFR (0.922 vs. 0.896) and PSSZ (0.934 vs. 0.895), with  $p < 10^{-5}$  based on DeLong's test. Similarly, non-Hispanic participants outperformed Hispanic participants (PSFR: 0.931 vs. 0.905; PSSZ: 0.965 vs. 0.909;  $p < 0.001$ ). These findings underscore the influence of demographic composition on risk model performance and reinforce the value of demographic alignment. In our study, the DISTINCT algorithm was applied to construct demographically matched NLST subsamples, supporting unbiased cross-comparisons between real and virtual imaging datasets.