scientific reports

Check for updates

OPEN

Techniques to produce and evaluate realistic multivariate synthetic data

John Heine[1]✉, Erin E. E. Fowler[1], Anders Berglund[2], Michael J. Schell[2] & Steven Eschrich[2]

Data modeling requires a sufficient sample size for reproducibility. A small sample size can inhibit model evaluation. A synthetic data generation technique addressing this small sample size problem is evaluated: from the space of arbitrarily distributed samples, a subgroup (class) has a latent multivariate *normal characteristic*; synthetic data can be generated from this class with univariate kernel density estimation (KDE); and synthetic samples are statistically like their respective samples. Three samples (n = 667) were investigated with 10 input variables (X). KDE was used to augment the sample size in X. Maps produced univariate normal variables in Y. Principal component analysis in Y produced uncorrelated variables in T, where the probability density functions were approximated as normal and characterized; synthetic data was generated with normally distributed univariate random variables in T. Reversing each step produced synthetic data in Y and X. All samples were approximately multivariate normal in Y, permitting the generation of synthetic data. Probability density function and covariance comparisons showed similarity between samples and synthetic samples. A class of samples has a latent *normal characteristic*. For such samples, this approach offers a solution to the small sample size problem. Further studies are required to understand this latent class.

Data modeling requires sufficient data for exploration and reproducibility purposes. This is especially relevant to biomedical-healthcare research, where data can be limited; although this field is broad, a few examples include risk prediction[1], response to therapy[2] and benign malignant classification[3–5]. Unfortunately, data can be limited for variety of reasons: the study of low-incidence diseases or underserved/underrepresented subpopulations[6]; clinic visitation hesitancy[7,8]; the inability to share data across facilities[9]; cost of molecular tests; and study timeframes. We, the authors, have worked in biomedical-healthcare research for many years and have experienced this persistent problem over decades.

Multivariate modeling is often exploratory that can decrease model stability in various ways. Here we explain frequent approaches that we have experienced or witnessed. The process starts by analyzing data from the target population for a variety of goals such as: open-ended analyses by studying many different data characteristics searching for correlations and patterns; subgrouping the dataset; testing hypothesis feasibilities with varying endpoints; exploring multiple hypotheses simultaneously; feature selection; selecting the most suitable model; or estimating model parameters with an optimization procedure. In practice, there are virtually *unlimited* ways to search through a given data sample. Data mining of this sort may not always be viewed in the most positive light[10], but on the other hand it is also the nature of discovery, noting there is often a compromise between these positions. Extensive subgroup analyses can effectively deplete the sample. When this applies, we term it the *small sample problem*. In the *final* stage, the fully specified model (i.e., the model with its parameters fixed) is validated with new data to prove its generalizability. Both the exploration and final stages depend critically on having an adequate sample size.

Determining the adequate sample size in the multivariate setting is a difficult task[11] and has relevance to the small sample problem. Adequate multivariate sample size is a function of both the analysis technique and covariance structure. For example, a multivariate two-sample test with normally distributed data and common covariance, Hotelling's T^2, is appropriate when comparing mean vectors[12]. In a broad sense, when the variables under consideration tend to be more highly correlated, the adequate sample size decreases and vice versa. Adequate sample size is a function of the number of free model parameters, which does not necessarily correspond with the number of variables[13]. In ordinary linear regression modeling with d noninteracting variables, there are about d parameters that must be determined. In contrast when taken to the limit, partial least squares regression[14] has

[1]Cancer Epidemiology Department, Moffitt Cancer Center and Research Institute, 12902 Bruce B. Downs Blvd, Tampa, FL 33612, USA. [2]Department of Biostatistics and Bioinformatics, Moffitt Cancer Center and Research Institute, 12902 Bruce B. Downs Blvd, Tampa, FL 33612, USA. ✉email: john.heine@moffitt.org

roughly d^2 parameters, and deep neural network architectures have even greater number of parameters, requiring large sample sizes for a given design[15] (see related table in[15] for examples). These modeling techniques illustrate that an adequate sample size under one condition may not be optimal for another. It is our premise that the adequate sample size for a given multivariate prediction problem that allows independent validation deserves more attention beyond *larger sample sizes are better,* especially when normality assumptions do not hold. By hypothesis, a technique that can generate realistic synthetic data will provide benefits in modeling endeavors. Such an approach could be used to augment an inadequate sample size for modeling and validation purposes or to study sample size requirements for a given multivariate covariance structure.

Synthetic data applications in health-related research use a variety of techniques. Some methods are used for generating samples from large populations[16–19]. These approaches include hidden Markov models[18], techniques that reconstruct time series data coupled with sampling the empirical probability density function of the relevant variables[16], and methods that estimate probability density functions (pdfs) from the data, not accounting for variable correlation[19]. Other work used moment matching to generate synthetic data but does not consider relative frequencies in the comparison analysis[20]. Discussions on synthetic data generation techniques indicate that the small sample size condition has received little analytical attention[20,21].

Our synthetic data technique estimates a multivariate pdf for arbitrarily distributed data, including when normal approximations fail to hold[22]. This initial work, based on multivariate kernel density estimation (mKDE) with unconstrained bandwidths, was illustrated with $d = 5$ data from mammographic case–control data[22]. Synthetic populations (SPs) of arbitrarily large size were generated from samples of limited size. However, mKDE has noted efficiency problems for high dimensionality[23,24]. As the dimensionality increases, the sample size requirement apparently becomes exceeding large, noting this area is under investigation. Although categorizing a problem as high or low dimensionality may be dependent on many factors, reasonable arguments suggest that *high dimensionality* may be defined as $3 < d \leq 50$, where d is the number of variables considered[25]; in that, density estimators should be able to address this range[25]. Here we let $d = 10$ so that many of the findings can be presented graphically or reasonably tabulated, and modeling problems in healthcare research can be within this range.

In this report, we present modifications to our method to mitigate the mKDE efficiency problem under specific conditions (latent normality) and address synthetic data generation in relatively higher dimensionality ($d = 10$). This modified approach decomposes an arbitrarily distributed multivariate problem into multiple univariate KDE (uKDE) problems while characterizing the covariance structure independently[26]. We are evaluating whether this approach can transform an arbitrarily distributed multivariate sample into an approximate multivariate normal form, which we define as a sample with a *latent normal characteristic*. In the universe of arbitrarily distributed samples, there is a multivariate normal subgroup. The technique for generating synthetic data for this normal subgroup is relatively straightforward and well-practiced. By hypothesis, our approach seeks to extend these straightforward techniques to the latent normal class by determining when (or if) it exists. Developing the analytics to detect this condition and then leveraging it to generate synthetic data are essential elements of our work[26].

Methods

Overview. Our modified synthetic data generation and analytic techniques have sequential components and many related analyses. Therefore, clear definitions, preliminaries, and a brief outline are given before the details are provided. Justifications are also discussed here when warranted.

Definitions. *Population* is used to define a hypothetical collection of virtually *unlimited* number of either real or synthetic entities from which samples comprised of observations or realizations may exist or can be drawn. The exception for the use of *population* is when explaining differential evolution (DE) optimization[27] used for uKDE bandwidth determination. The *DE-population* is limited and defined specifically. *Sample* defines a collection of n real observations with d attributes (variables) from the space of possible samples, represented mathematically as a $n \times d$ matrix (rows = observations, columns = attributes). Column vectors are designated with lower-case bold letters. For example, individual attributes are referred to as $\mathbf{x}$, a column vector. Vector components are designated with lower-case subscripted letters. The components of $\mathbf{x}$ are referenced as x_j for $j = 1, 2, \ldots, d$ and assumed continuous. The multivariate pdf for $\mathbf{x}$ is $p(\mathbf{x})$. X refers to the input variable space, that is, $\mathbf{x}$ exists in the X representation. We assume $p(\mathbf{x})$ exists at the population level, but not accessible. In practice, we evaluated normalized histograms throughout this work for all variables considered both univariate and multivariate (i.e., empirical pdfs), also referred to as pdfs for brevity; we use this term to refer to attributes at the population level as well as at the sample level. One-dimensional (1D) marginal pdfs for $p(\mathbf{x})$ are expressed as $p_j(x_j)$. Matrices are designated with upper case bold lettering. For example, $\mathbf{X}$ is the $n \times d$ matrix with n observations of $\mathbf{x}$ in its rows (i.e., the ith row contains the d attributes for the ith observation and the jth column of $\mathbf{x}$ has n realizations of x_j). Double subscripts are used for both specific realizations and matrix element indices. That is, x_{ij} is the jth component for the ith realization in X (also is the indexing for $\mathbf{X}$). Variables in X are mapped to the Y representation. This creates the corresponding entities in Y: (1) the vector $\mathbf{y}$ with d components; (2) the multivariate pdf, $g(\mathbf{y})$, and its marginal pdfs, $g_j(y_j)$; and (3) the matrix $\mathbf{Y}$ defined analogously as $\mathbf{X}$. We also work in the T representation (uncorrelated variables) as explained below, where $\mathbf{t}$, t_j, and $\mathbf{T}$ are defined similarly. Likewise, $r(\mathbf{t})$ is the multivariate pdf in T with marginals, $r_j(t_j)$. We define the cumulative probability functions (i.e., the indefinite integral approximation of a given univariate pdf) for $p_j(x_j)$ and $g_j(y_j)$ as $P_j(x_j)$ and $G_j(y_j)$, respectively. Covariance quantities are calculated with the normal multivariate form: $E(w - m_w)(v - m_v)$, where E is the expectation operator, w and v are arbitrary random variables with means m_w and m_v. The corresponding covariance matrices are expressed as $\mathbf{C}_x$, $\mathbf{C}_y$, and $\mathbf{C}_t$, respectively (or $\mathbf{C}_k$ generically). When an entity is given the subscript, s, it then defines the corresponding synthetic entity. Standardized normal defines a zero mean–unit variance normal pdf,

used in both the univariate and multivariate scenarios. *Parametric* for this report refers to functions that can be expressed in closed form.

Preliminaries. General biomedical healthcare data characteristics are discussed to overview some of their characteristics. Measurements such as body mass index (BMI) and age, or measurements taken from image data can have right-skewed pdfs because they are often positive-valued and not inclusive of zero (see Figs. 1, 2 and 3 for examples). Such measures can bear varying levels of correlation (see lower parts of Tables 1, 2 and 3). Thus, an arbitrary p(**x**) may not lend itself to parametric modeling in X (i.e., normality and non-correlation assumptions do not apply in many instances). To render such data into a more tractable form, a series of steps (see Fig. 4) were taken to condition **X**; by premise, these steps will permit characterizing the sample with parametric means and then generating similar multivariate synthetic data without mKDE.

Outline of the processing steps. When describing these steps, we also briefly discuss the analysis at a given step (also provided in detail in the methods). The process starts with a given sample in X (Fig. 4, top left). The processing flow for the sample (X–Y–T) is illustrated in the top row of Fig. 4, and the reversed SP generation flow (T_s–Y_s–X_s) in the bottom row.

Step 1 Univariate maps were constructed (Fig. 4, top-left) to transform a given X measurement to a standardized normal, producing the respective marginal pdf set in Y (Fig. 4, top-middle). Maps were constructed with

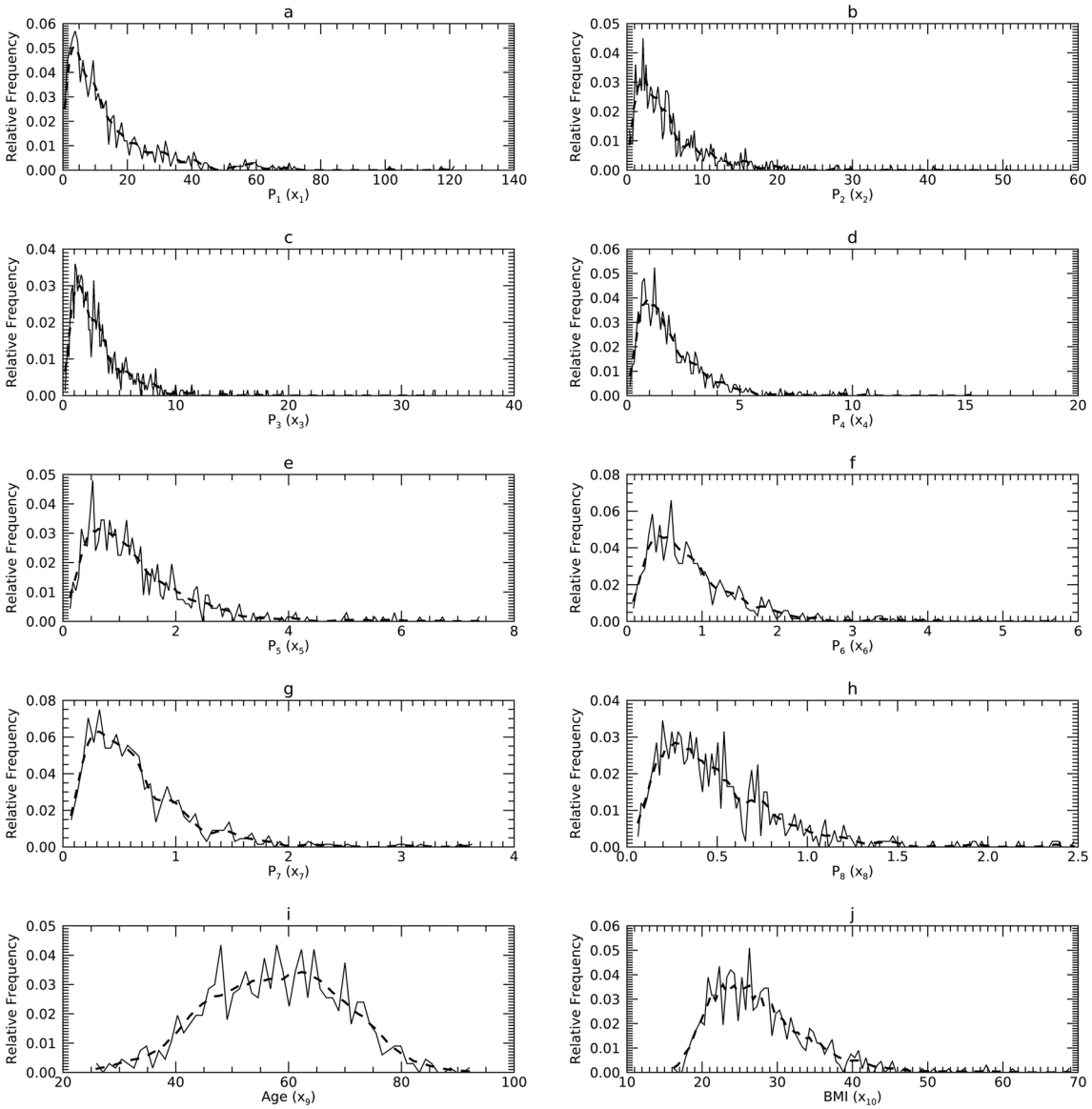

Figure 1. Marginal Probability Density Functions (pdfs) for Sample 1 (DS1) in the X representation: each pdf for DS1 (solid) is compared with its corresponding pdf from synthetic data (dashes). The x-axis cites the variable name from its respective resource and its index name parenthetically (x_j).

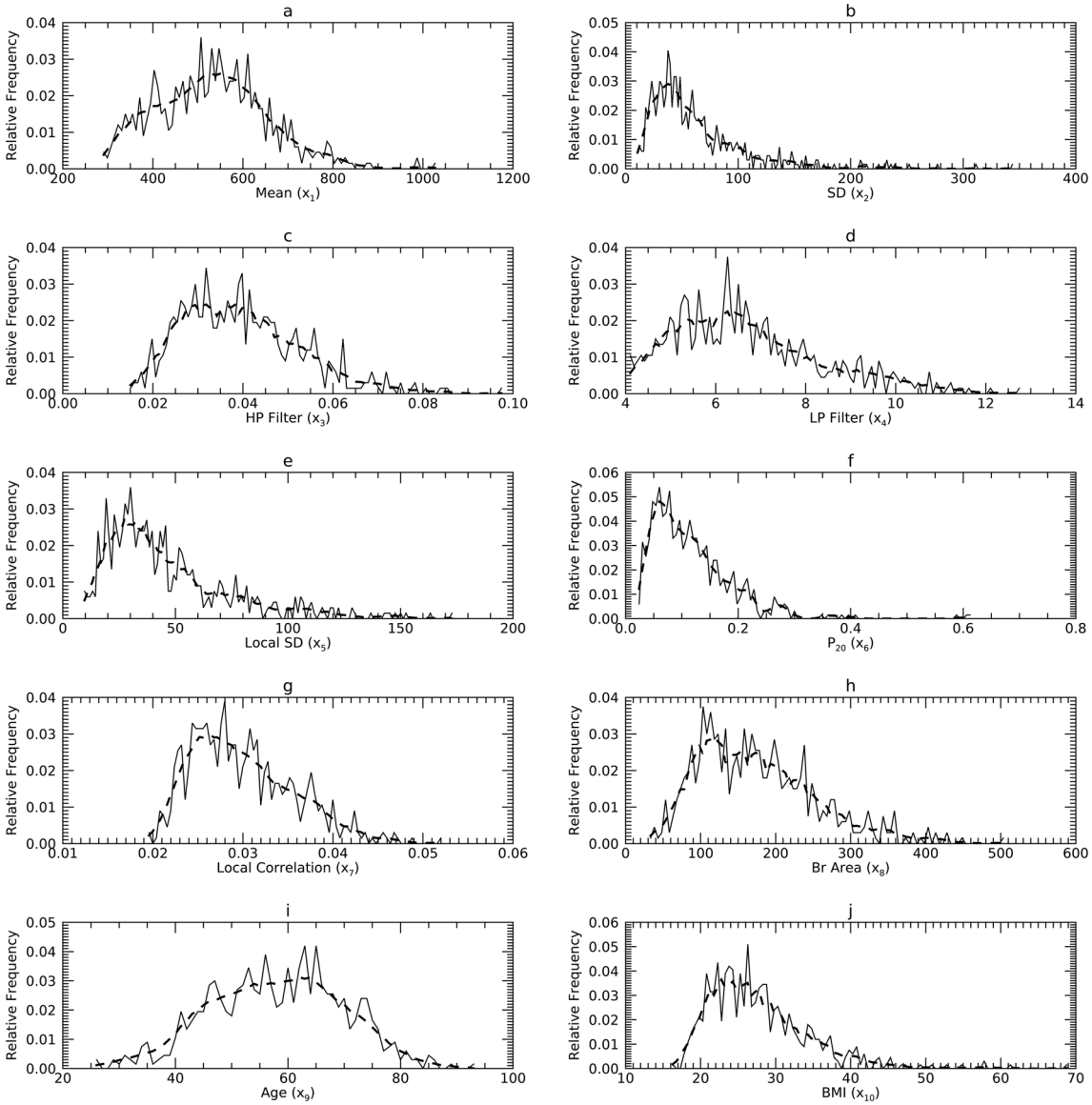

Figure 2. Marginal Probability Density Functions (pdfs) for Sample 2 (DS2) in the X representation: each pdf for DS2 (solid) is compared with its corresponding pdf from synthetic data (dashes). The x-axis cites the variable name from its respective resource and its index name parenthetically (x_j).

an augmented sample size using optimized uKDE, addressing the small sample size problem. uKDE was used to generate synthetic x_j. Here, we augmented the sample size with the goal of filling gaps in the input marginal pdfs of x_j (sample) to complement the map constructions. By hypothesis, this step addresses the small sample size problem by guaranteeing continuous smooth maps that will produce standardized normal pdfs from the sample. Synthetic x_j generated in this fashion do not maintain the covariance relationships in X and were not used further; only x_j from the sample were mapped to Y, and KDE was not used beyond this point. There is no guarantee that a set of normal marginals in Y will produce a multivariate normal pdf. Although the reverse is always true because a multivariate normal has univariate normal marginals. In practice, $g(\mathbf{y})$ from the sample could be assessed at this point to estimate how well it approximates normality. If the latent normal approximation is poor, another synthetic approach could be pursued, or the process could be discontinued. Here we forgo such testing at this step (normality was tested for in steps 3 and 4 instead) and move through all steps to illustrate the techniques. We will also discuss a possible modification that could be investigated when the sample has a poor latent normal characteristic approximation, later in the discussion.

Step 2 Principal component analysis (PCA) was used to decouple the variables in Y producing uncorrelated variables in T (Fig. 4, top-right).

Step 3 Synthetic data was generated in T (Fig. 4, bottom-left) as uncorrelated random variables. Here we assumed that each marginal in T from the sample could be approximated as normal with variance given by the jth eigenvalue of $\mathbf{C}_y$. To generate synthetic data, the columns of $\mathbf{T}_s$ were populated as normally distributed random variables with these specified variances (noting, the columns lack correlation). We refer to the realizations in T as the SP (i.e., $\mathbf{T}_s$), noting the column length (number of synthetic entities) can be arbitrarily large.

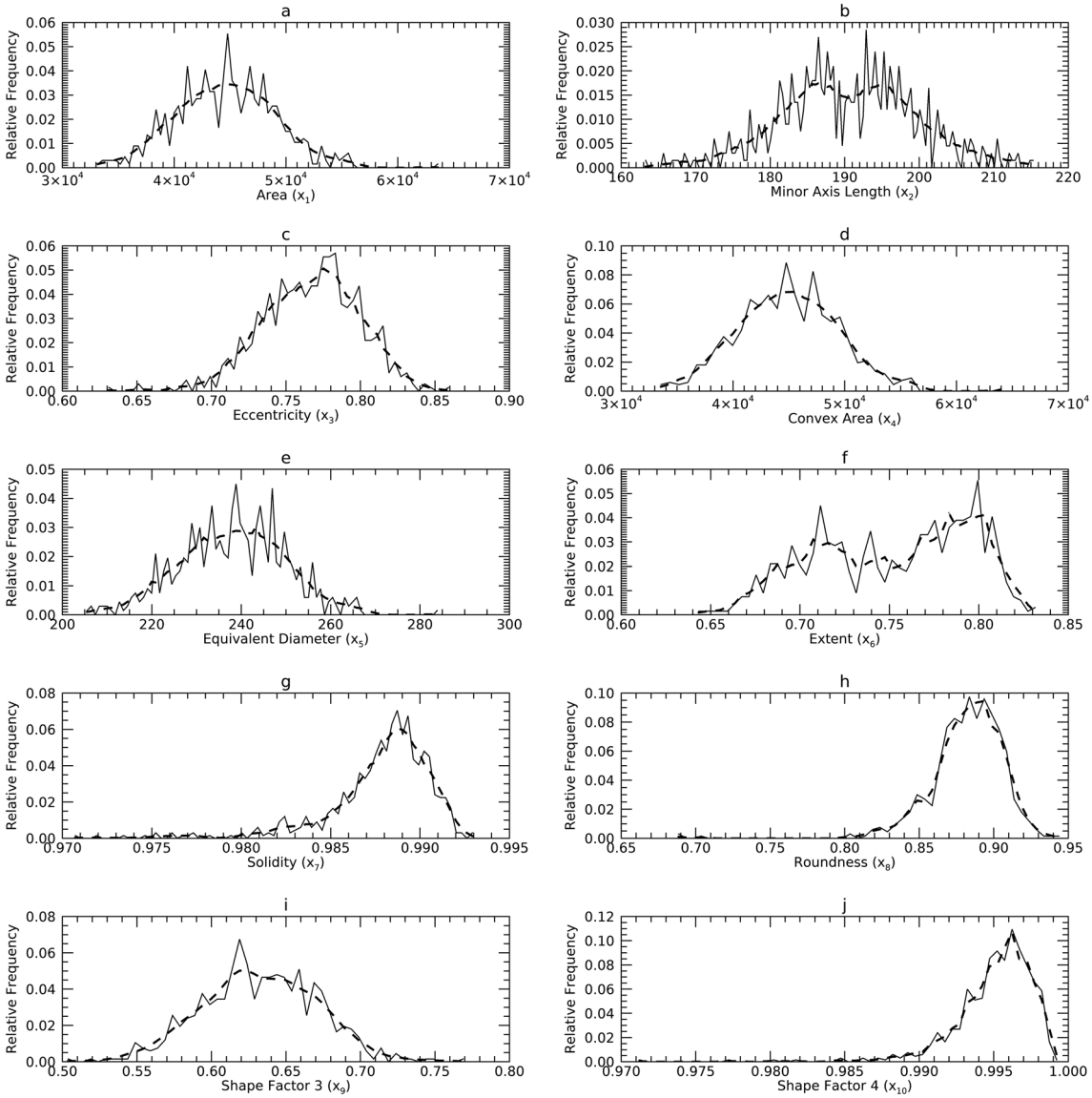

Figure 3. Marginal Probability Density Functions (pdfs) for Sample 3 (DS3) in the X representation: each pdf for DS3 (solid) is compared with its corresponding pdf from synthetic data (dashes). The x-axis cites the variable name from its respective dataset and its index name parenthetically (x_j).

To address the normal characteristic, univariate marginal and multivariate pdfs from the sample were tested for normality at this point.

Step 4 The inverse PCA transform (Fig. 4, bottom-middle) of $\mathbf{T}_s$ produced the SP in Y ($\mathbf{Y}_s$), thereby restoring the covariance relationships that were removed in Step 2. We note, this technique (steps 3 and 4) of producing multivariate normal data is a practiced approach when the sample is multivariate normal or well approximated as such. For reference, the multivariate standardized normal in Y is expressed as

$$g_n(\mathbf{y}) = (2\pi)^{-\frac{d}{2}} |\mathbf{C}_y|^{-\frac{1}{2}} \exp - \frac{1}{2}\left[\mathbf{y}^T \mathbf{C}_y^{-1} \mathbf{y}\right], \quad (1)$$

where $\mathbf{C}_y$ can be approximated as the covariance matrix from a given sample. If synthetic samples are poor replicas of their sample, it follows the sample's g(y) will be a poor approximation of multivariate normality (i.e., the sample is not in the latent normal class).

We evaluated how well the inverse PCA transformation preserved the covariance ($\mathbf{C}_y$) and the method's capability of restoring the univariate/multivariate pdfs in Y (i.e., normality comparisons) rather than in Step 2.

Step 5 Each synthetic variable in Y is inverse mapped to X (Fig. 4, bottom left). This reversed Step 1, thereby producing the SP in X ($\mathbf{X}_s$), and restoring the covariance relationships by hypothesis. The respective pdfs and covariance matrices in X were compared with those from synthetic samples; pdfs were also tested for normality.

It is important to clarify a few aspects of this work. The univariate mapping from X to Y creates a set of univariate marginals normally distributed that can produce a multivariate normal, but not guaranteed. The

a

	P_1 (x_1)	P_2 (x_2)	P_3 (x_3)	P_4 (x_4)	P_5 (x_5)	P_6 (x_6)	P_7 (x_7)	P_8 (x_8)	Age (x_9)	BMI (x_{10})
P_1 (x_1)	**2.27E+02**	7.27E+01	3.61E+01	1.97E+01	1.18E+01	8.16E+00	5.76E+00	4.27E+00	− 4.13E+01	− 1.98E+01
	(1.69E+02, 2.88E+02)	(5.38E+01, 9.44E+01)	(2.54E+01, 4.83E+01)	(1.47E+01, 2.58E+01)	(9.17E+00, 1.53E+01)	(6.16E+00, 1.04E+01)	(4.30E+00, 7.19E+00)	(3.21E+00, 5.26E+00)	(− 5.15E+01, − 2.20E+01)	(− 2.59E+01, − 1.18E+01)
P_2 (x_2)	0.9150	**2.78E+01**	1.41E+01	7.78E+00	4.68E+00	3.21E+00	2.25E+00	1.66E+00	− 9.34E+00	− 5.59E+00
		(1.97E+01, 3.62E+01)	(9.42E+00, 1.89E+01)	(5.55E+00, 1.01E+01)	(3.52E+00, 5.96E+00)	(2.35E+00, 4.06E+00)	(1.64E+00, 2.81E+00)	(1.23E+00, 2.05E+00)	(− 1.49E+01, − 4.84E+00)	(− 7.08E+00, − 2.12E+00)
P_3 (x_3)	0.8570	0.9600	**7.82E+00**	4.30E+00	2.58E+00	1.77E+00	1.22E+00	8.95E−01	− 4.22E+00	− 2.29E+00
			(4.78E+00, 1.10E+01)	(2.89E+00, 5.73E+00)	(1.87E+00, 3.35E+00)	(1.25E+00, 2.31E+00)	(8.75E−01, 1.59E+00)	(6.61E−01, 1.16E+00)	(− 6.83E+00, − 1.60E+00)	(− 2.92E+00, − 2.31E−01)
P_4 (x_4)	0.8230	0.9270	0.9660	**2.53E+00**	1.52E+00	1.04E+00	7.18E−01	5.32E−01	− 1.65E+00	− 1.33E+00
				(1.81E+00, 3.28E+00)	(1.15E+00, 1.93E+00)	(7.74E−01, 1.32E+00)	(5.41E−01, 9.14E−01)	(4.09E−01, 6.73E−01)	(− 3.25E+00, − 3.05E−01)	(− 1.63E+00, − 9.54E−02)
P_5 (x_5)	0.8020	0.9060	0.9410	0.9760	**9.59E−01**	6.48E−01	4.47E−01	3.32E−01	− 9.87E−01	− 8.59E−01
					(7.51E−01, 1.20E+00)	(5.03E−01, 8.07E−01)	(3.52E−01, 5.63E−01)	(2.65E−01, 4.17E−01)	(− 1.96E+00, − 1.37E−01)	(− 1.03E+00, − 6.19E−02)
P_6 (x_6)	0.8030	0.9040	0.9380	0.9720	0.9810	**4.54E−01**	3.12E−01	2.33E−01	− 6.05E−01	− 5.83E−01
						(3.49E−01, 5.67E−01)	(2.42E−01, 3.90E−01)	(1.84E−01, 2.89E−01)	(− 1.23E+00, 2.02E−02)	(− 6.93E−01, − 2.55E−02)
P_7 (x_7)	0.8100	0.9040	0.9230	0.9570	0.9690	0.9820	**2.22E−01**	1.65E−01	− 4.09E−01	− 4.28E−01
							(1.73E−01, 2.78E−01)	(1.30E−01, 2.04E−01)	(− 8.82E−01, − 8.18E−03)	(− 4.93E−01, − 2.36E−02)
P_8 (x_8)	0.8000	0.8900	0.9050	0.9450	0.9590	0.9750	0.9900	**1.25E−01**	− 2.62E−01	− 3.10E−01
								(9.99E−02, 1.54E−01)	(− 5.91E−01, 5.92E−02)	(− 3.64E−01, − 1.00E−02)
Age (x_9)	− 0.2330	− 0.1510	− 0.1290	− 0.0884	− 0.0859	− 0.0765	− 0.0739	− 0.0631	**1.38E+02**	− 4.74E+00
									(1.30E+02, 1.57E+02)	(− 1.09E+01, 7.87E−01)
BMI (x_{10})	− 0.1950	− 0.1580	− 0.1220	− 0.1240	− 0.1300	− 0.1290	− 0.1350	− 0.1300	− 0.0600	**4.52E+01**
										(3.72E+01, 5.31E+01)

b

	P_1 (y_1)	P_2 (y_2)	P_3 (y_3)	P_4 (y_4)	P_5 (y_5)	P_6 (y_6)	P_7 (y_7)	P_8 (y_8)	Age (y_9)	BMI (y_{10})
P_1 (y_1)	**1.00E+00**	9.48E−01	9.07E−01	8.78E−01	8.59E−01	8.49E−01	8.43E−01	8.31E−01	− 2.35E−01	− 2.29E−01
	(9.03e−01, 1.12e+00)	(8.52e−01, 1.06e+00)	(8.11e−01, 1.02e+00)	(7.85e−01, 9.86e−01)	(7.67e−01, 9.66e−01)	(7.54e−01, 9.56e−01)	(7.47e−01, 9.46e−01)	(7.34e−01, 9.35e−01)	(− 3.18e−01, − 1.60e−01)	(− 3.08e−01, − 1.53e−01)
P_2 (y_2)	0.9480	**1.00E+00**	9.71E−01	9.49E−01	9.34E−01	9.22E−01	9.16E−01	9.03E−01	− 1.78E−01	− 1.54E−01
		(9.01e−01, 1.11e+00)	(8.70e−01, 1.09e+00)	(8.52e−01, 1.06e+00)	(8.38e−01, 1.05e+00)	(8.25e−01, 1.03e+00)	(8.17e−01, 1.03e+00)	(8.08e−01, 1.01e+00)	(− 2.56e−01, − 1.02e−01)	(− 2.33e−01, − 8.02e−02)
P_3 (y_3)	0.9070	0.9710	**1.00E+00**	9.78E−01	9.68E−01	9.61E−01	9.54E−01	9.43E−01	− 1.45E−01	− 1.00E−01
			(9.01e−01, 1.12e+00)	(8.79e−01, 1.09e+00)	(8.69e−01, 1.08e+00)	(8.63e−01, 1.07e+00)	(8.55e−01, 1.07e+00)	(8.45e−01, 1.05e+00)	(− 2.26e−01, − 6.76e−02)	(− 1.77e−01, − 2.53e−02)
P_4 (y_4)	0.8780	0.9490	0.9780	**1.00E+00**	9.85E−01	9.79E−01	9.74E−01	9.65E−01	− 1.05E−01	− 9.38E−02
				(9.01e−01, 1.11e+00)	(8.86e−01, 1.10e+00)	(8.82e−01, 1.09e+00)	(8.75e−01, 1.09e+00)	(8.67e−01, 1.08e+00)	(− 1.85e−01, − 2.75e−02)	(− 1.73e−01, − 1.72e−02)
P_5 (y_5)	0.8590	0.9340	0.9680	0.9850	**1.00E+00**	9.86E−01	9.82E−01	9.74E−01	− 9.71E−02	− 9.27E−02
					(9.00e−01, 1.11e+00)	(8.87e−01, 1.10e+00)	(8.81e−01, 1.09e+00)	(8.73e−01, 1.08e+00)	(− 1.76e−01, − 2.39e−02)	(− 1.73e−01, − 1.88e−02)
P_6 (y_6)	0.8490	0.9220	0.9610	0.9790	0.9860	**1.00E+00**	9.89E−01	9.82E−01	− 8.20E−02	− 8.90E−02
						(9.01e−01, 1.11e+00)	(8.89e−01, 1.10e+00)	(8.80e−01, 1.10e+00)	(− 1.62e−01, − 7.43e−03)	(− 1.72e−01, − 1.48e−02)
P_7 (y_7)	0.8430	0.9160	0.9540	0.9740	0.9820	0.9890	**1.00E+00**	9.89E−01	− 8.48E−02	− 9.07E−02
							(8.95e−01, 1.11e+00)	(8.83e−01, 1.10e+00)	(− 1.65e−01, − 1.01e−02)	(− 1.72e−01, − 1.52e−02)
P_8 (y_8)	0.8310	0.9030	0.9430	0.9650	0.9740	0.9820	0.9890	**1.00E+00**	− 6.30E−02	− 8.84E−02
								(8.96e−01, 1.11e+00)	(− 1.43e−01, 8.14e−03)	(− 1.68e−01, − 9.24e−03)

Continued

b										
	P_1 (y_1)	P_2 (y_2)	P_3 (y_3)	P_4 (y_4)	P_5 (y_5)	P_6 (y_6)	P_7 (y_7)	P_8 (y_8)	Age (y_9)	BMI (y_{10})
Age (y_9)	− 0.2350	− 0.1780	− 0.1450	− 0.1050	− 0.0971	− 0.0820	− 0.0848	− 0.0630	**1.00E+00**	− 6.64E−02
									(8.88e−01, 1.11e+00)	(− 1.44e−01, 1.88e−02)
BMI (y_{10})	− 0.2290	− 0.1540	− 0.1000	− 0.0938	− 0.0927	− 0.0890	− 0.0907	− 0.0884	− 0.0664	**1.00E+00**
										(8.87e−01, 1.11e+00)

Table 1. Covariance and correlation for Dataset 1: in both tables, entries on the diagonals and above give covariance quantities. Entries below the diagonals (bold) provide the respective Pearson correlation coefficients. Table 1a gives the X representation quantities and 1b the Y representation quantities. The covariance quantities were generated from the respective sample. Parenthetically, 95% confidence intervals generated from synthetic samples are cited below the respective covariance quantity. Variables are cited with the names used in their respective resource and with the names used in this report parenthetically.

comparison of univariate marginal pdfs, however, is no guarantee that the respective multivariate pdfs are reasonable facsimiles because the covariance structure has been removed. Many univariate pdf comparisons are provided between samples and synthetic samples in addition to multivariate comparisons, because these allow visualizing similarities with the above stipulations. When a given sample has the latent normal characteristic, the SP generation is greatly simplified, and then it is *defined* by Eq. (1). The work below shows how to generate synthetic data when this characteristic holds. We use three datasets that were selected *pseudo-randomly*. In the space of samples (virtually unlimited), we do not know the probability that a sample selected at random will have this latent normal characteristic. The main objectives are to present the analysis components with the methods for testing for the latent characteristic, give a thorough investigation, demonstrate that the synthesis produces realistic data when this latent condition exists, and then discuss further analyses.

Study data. Samples were derived from two sources of measurements: (1) mammograms and related clinical data (n = 667), and (2) dried beans (n = 13,611)[28]. Most technical aspects of these data are not relevant for this report. Mammography data included all observations with mammograms from a specific imaging technology, thereby defining n = 667. We used the dried bean data to add variation to the analyses as the variable nomenclatures are very different from the mammogram data, noting at this point the source of data is not germane. From mammograms, we considered two sets of measurements each with d = 10 variables referred to as Sample 1 (DS1) and Sample 2 (DS2). DS1 has 8 double precision measurements from the Fourier power spectrum in addition to age and BMI, both captured as integer variables. The Fourier attributes are from a set of measurements described previously[29]; the first 8 measurements from this set are labeled as P_i for i = 1–8. These Fourier measures are consecutive and follow an approximate functional form[30], and thus represent variables that are very different than those in DS2 (or Sample 3 below). To cite the covariance quantities and correlation coefficients, we used a modified covariance (covariance for short) table format for efficiency because $\mathbf{C}_k$ is symmetric. In these tables, entries below the diagonal are the respective correlation coefficients, whereas the elements along the diagonal (variance quantities) and above are the covariance quantities. The covariance table for DS1 is shown in Table 1a. DS2 contains 8 double precision summary measurements derived from the image domain: mean, standard deviation (SD); SD of a high-pass (HP) filter output, SD of a low-pass (LP) filter output; local SD summarized; P_{20} Fourier measure (from the set described for DS1 measurements); local spatial correlation summarized[31]; and breast (Br) area measured in cm^2. Age and BMI (from DS1) were also included in this dataset. These variables were selected virtually at random to give d = 10 and possibly provide a different covariance structure than DS1. The covariance table is shown in Table 2a. Neither DS1 nor DS2 were used in our related-prior mKDE synthetic data work. Selected measures and realizations from the dried bean dataset[28] are referred to as Sample 3 (DS3). The bean data has 17 measurements (floating point) from 7 bean types. We selected 10 measures at random to make the dimensionality of DS3 compatible with the other two datasets giving this set of variables: area (1); minor axis (5); eccentricity (6); convex area (7); equivalent diameter (8); extent (9); solidarity (10); roundness (11); shape factor 3 (15), and shape factor 4 (16). Here, parenthetical references give the variable number listed in the respective resource (see[28]). Both bean type (bean type = Sira, with n = 2636) and n = 667 observations were selected at random to create DS3. Keeping n = 667 constant across datasets permits consistent statistical comparisons. For example, confidence intervals (CIs) and other comparison metrics are dependent upon the number of observations. The covariance table is shown in Table 3a. The analysis of three samples supports the evaluation of the processing scheme under generalized scenarios. The means and standard deviations for x_j in each dataset are provided in Table 4. Note, the dynamic range of the means and standard deviations within a given sample vary widely in some instances.

KDE, optimization, and mapping. The mapping (Step 1) relies on generating synthetic x_j with uKDE. Each bandwidth parameter was determined with an optimization process wherein synthetic data from uKDE was compared with the sample. There is a continued feedback loop between the sample, synthetic data generation, and comparison during the optimization process. When the optimization was completed, a given map was constructed.

a

	Mean (x_1)	SD (x_2)	HP filter (x_3)	LP filter (x_4)	Local SD (x_5)	P_{20} (x_6)	Local correlation (x_7)	Br area (x_8)	Age (x_9)	BMI (x_{10})
Mean (x_1)	**1.53E+04**	2.80E+03	1.23E+00	1.73E+02	2.02E+03	7.09E+00	−8.77E−02	1.20E+03	−2.30E+02	1.29E+02
	(1.37e+04, 1.71e+04)	(2.43e+03, 3.54e+03)	(1.07e+00, 1.41e+00)	(1.57e+02, 1.99e+02)	(1.70e+03, 2.36e+03)	(6.09e+00, 8.35e+00)	(−1.38e−01, −3.02e−02)	(7.05e+02, 2.28e+03)	(−3.31e+02, −1.03e+02)	(9.72e+01, 2.24e+02)
SD (x_2)	0.5170	**1.91E+03**	4.40E−01	2.55E+01	1.05E+03	1.40E+00	−1.07E−01	−4.02E+02	−8.94E+01	−3.38E+01
		(1.50e+03, 2.28e+03)	(3.71e−01, 5.21e−01)	(2.55e+01, 3.94e+01)	(9.20e+02, 1.32e+03)	(1.32e+00, 2.14e+00)	(−1.39e−01, −9.85e−02)	(−6.93e+02, −2.05e+02)	(−1.56e+02, −7.61e+01)	(−4.56e+01, −2.62e+00)
HP filter (x_3)	0.7500	0.7610	**1.75E−04**	1.46E−02	2.74E−01	7.72E−04	−3.54E−05	−2.17E−01	−1.50E−02	−1.37E−02
			(1.57e−04, 1.98e−04)	(1.30e−02, 1.73e−02)	(2.42e−01, 3.27e−01)	(6.64e−04, 9.15e−04)	(−4.16e−05, −2.91e−05)	(−2.64e−01, −1.08e−01)	(−2.80e−02, −2.94e−03)	(−1.60e−02, −3.26e−03)
LP filter (x_4)	0.8540	0.3560	0.6760	**2.68E+00**	1.97E+01	1.01E−01	4.61E−04	2.00E+01	−2.59E+00	1.47E+00
				(2.36e+00, 2.97e+00)	(1.79e+01, 2.61e+01)	(8.51e−02, 1.17e−01)	(−3.44e−04, 1.08e−03)	(1.40e+01, 3.57e+01)	(−4.11e+00, −1.20e+00)	(1.21e+00, 2.99e+00)
Local SD (x_5)	0.5890	0.8670	0.7460	0.4340	**7.70E+02**	1.15E+00	−9.14E−02	−4.69E+02	−8.12E+01	−3.36E+01
					(6.53e+02, 8.85e+02)	(9.51e−01, 1.44e+00)	(−1.04e−01, −7.88e−02)	(−5.89e+02, −2.84e+02)	(−1.12e+02, −6.19e+01)	(−3.89e+01, −1.15e+01)
P_{20} (x_6)	0.7750	0.4340	0.7910	0.8320	0.5630	**5.46E−03**	−8.46E−05	−2.26E−01	−4.60E−02	−1.95E−02
						(4.15e−03, 6.39e−03)	(−1.08e−04, −4.57e−05)	(−4.08e−01, 4.41e−01)	(−1.27e−01, 8.43e−03)	(−1.87e−02, 5.41e−02)
Local correlation (x_7)	−0.1250	−0.4310	−0.4720	0.0497	−0.5800	−0.2020	**3.22E−05**	2.41E−01	1.58E−02	1.56E−02
							(2.93e−05, 3.65e−05)	(1.99e−01, 2.82e−01)	(1.11e−02, 2.21e−02)	(1.20e−02, 1.91e−02)
Br area (x_8)	0.1230	−0.1170	−0.2090	0.1560	−0.2160	−0.0390	0.5420	**6.15E+03**	5.26E+00	3.25E+02
								(5.50e+03, 7.01e+03)	(−7.34e+01, 6.33e+01)	(2.84e+02, 4.00e+02)
Age (x_9)	−0.1580	−0.1740	−0.0966	−0.1350	−0.2490	−0.0530	0.2380	0.0057	**1.38E+02**	−4.74E+00
									(1.30e+02, 1.57e+02)	(−1.17e+01, 8.82e−01)
BMI (x_{10})	0.1550	−0.1150	−0.1540	0.1330	−0.1800	−0.0393	0.4070	0.6170	−0.0600	**4.52E+01**
										(3.81e+01, 5.34e+01)

b

	Mean (y_1)	SD (y_2)	HP filter (y_3)	LP filter (y_4)	Local SD (y_5)	P_{20} (y_6)	Local correlation (y_7)	Br area (y_8)	Age (y_9)	BMI (y_{10})
Mean (y_1)	**1.00E+00**	6.03E−01	7.56E−01	8.86E−01	6.24E−01	8.39E−01	−1.25E−01	1.53E−01	−1.48E−01	1.97E−01
	(8.96e−01, 1.11e+00)	(5.18e−01, 6.83e−01)	(6.62e−01, 8.46e−01)	(7.85e−01, 9.91e−01)	(5.32e−01, 7.04e−01)	(7.39e−01, 9.37e−01)	(−1.96e−01, −4.65e−02)	(7.79e−02, 2.34e−01)	(−2.24e−01, −7.16e−02)	(1.21e−01, 2.73e−01)
SD (y_2)	0.6030	**1.00E+00**	8.14E−01	4.90E−01	9.47E−01	5.83E−01	−5.64E−01	−1.49E−01	−2.46E−01	−9.86E−02
		(8.95e−01, 1.10e+00)	(7.20e−01, 9.08e−01)	(4.06e−01, 5.70e−01)	(8.41e−01, 1.05e+00)	(5.00e−01, 6.68e−01)	(−6.50e−01, −4.70e−01)	(−2.24e−01, −7.63e−02)	(−3.24e−01, −1.68e−01)	(−1.74e−01, −2.62e−02)
HP filter (y_3)	0.7560	0.8140	**1.00E+00**	7.05E−01	7.96E−01	8.44E−01	−4.89E−01	−1.86E−01	−9.96E−02	−1.14E−01
			(8.89e−01, 1.10e+00)	(6.09e−01, 7.94e−01)	(6.99e−01, 8.87e−01)	(7.44e−01, 9.43e−01)	(−5.75e−01, −3.98e−01)	(−2.61e−01, −1.09e−01)	(−1.77e−01, −2.22e−02)	(−1.87e−01, −4.23e−02)
LP filter (y_4)	0.8860	0.4900	0.7050	**1.00E+00**	5.12E−01	8.84E−01	3.63E−02	2.00E−01	−1.43E−01	2.01E−01
				(8.91e−01, 1.12e+00)	(4.25e−01, 5.94e−01)	(7.80e−01, 9.91e−01)	(−3.58e−02, 1.16e−01)	(1.21e−01, 2.85e−01)	(−2.13e−01, −6.31e−02)	(1.21e−01, 2.79e−01)
Local SD (y_5)	0.6240	0.9470	0.7960	0.5120	**1.00E+00**	6.27E−01	−6.51E−01	−2.18E−01	−2.79E−01	−1.52E−01
					(8.90e−01, 1.10e+00)	(5.38e−01, 7.11e−01)	(−7.43e−01, −5.54e−01)	(−2.96e−01, −1.42e−01)	(−3.55e−01, −2.01e−01)	(−2.33e−01, −7.82e−02)
P_{20} (y_6)	0.8390	0.5830	0.8440	0.8840	0.6270	**1.00E+00**	−2.07E−01	1.29E−03	−7.42E−02	3.86E−02
						(8.93e−01, 1.11e+00)	(−2.83e−01, −1.27e−01)	(−7.33e−02, 8.16e−02)	(−1.53e−01, 9.28e−04)	(−3.55e−02, 1.12e−01)
Local correlation (y_7)	−0.1250	−0.5640	−0.4890	0.0363	−0.6510	−0.2070	**1.00E+00**	5.36E−01	2.50E−01	4.22E−01
							(8.89e−01, 1.11e+00)	(4.50e−01, 6.33e−01)	(1.70e−01, 3.29e−01)	(3.35e−01, 5.08e−01)
Br area (y_8)	0.1530	−0.1490	−0.1860	0.2000	−0.2180	0.0013	0.5360	**1.00E+00**	−5.66E−03	6.57E−01
								(8.90e−01, 1.12e+00)	(−8.17e−02, 6.83e−02)	(5.63e−01, 7.54e−01)
Age (y_9)	−0.1480	−0.2460	−0.0996	−0.1430	−0.2790	−0.0742	0.2500	−0.0057	**1.00E+00**	−6.70E−02
									(9.00e−01, 1.11e+00)	(−1.48e−01, 7.79e−03)

Continued

b										
	Mean (y_1)	**SD (y_2)**	**HP filter (y_3)**	**LP filter (y_4)**	**Local SD (y_5)**	**P_{20} (y_6)**	**Local correlation (y_7)**	**Br area (y_8)**	**Age (y_9)**	**BMI (y_{10})**
BMI (y_{10})	0.1970	− 0.0986	− 0.1140	0.2010	− 0.1520	0.0386	0.4220	0.6570	− 0.0670	**1.00E+00**
										(8.91e−01, 1.11e+00)

Table 2. Covariance and correlation for Dataset 2: in both upper and lower tables, entries on the diagonals and above give covariance quantities. Entries below the diagonals (bold) provide the respective Pearson correlation coefficients. Table 2a gives the X representation quantities and 2b the Y representation quantities. The covariance quantities were generated from the respective sample. Parenthetically, 95% confidence intervals generated from synthetic samples are cited below the respective covariance quantity. Variables are cited with the names used in their respective resource and with the names used in this report parenthetically.

uKDE. For the map construction in Step 1 and as a modification to our previous work[22], uKDE was used to generate realizations from each $p_j(x_j)$ given by

$$p_j(x_j) = \frac{1}{k_j n}\sum_{i=i}^{n} \exp - \frac{(x_j - x_{ij})^2}{2h_j^2} = \frac{1}{n}\sum_{i=1}^{n} K(x_j), \tag{2}$$

where x_j is the synthetic variable for this discussion, x_{ij} are observations from a given sample, k_j is a normalization factor, and h_j is the univariate bandwidth parameter.

Optimization. Differential evolution optimization[27] was used to determine each h_j in Eq. (2). The population of candidate h_j evolves over generations to a *solution*, as described in detail previously[22]. To form generation zero for a given x_j, the DE-population (n = 1000) was initialized randomly (uniformly distributed) within these bounds: (0.0, 4 × the variance of x_j) because a given range should span the solution for h_j. The DE-population size stays constant across all generations. By expectation, the populations become more fit according to the fitness function over generations. A given generation is found by comparing two-DE populations (1000 pair-wise competitions) of candidate h_j solutions derived from the previous generation. For a given pairwise competition, two respective SPs were generated for each candidate h_j. Synthetic samples (n = 667) were drawn from each SP and $P_j(x_j)$ from the sample was compared with each $P_j(x_j)$ derived from its synthetic sample using Eq. (2). The h_j candidate [used in Eq. (2)] that produced a smaller D_j was used to populate the current generation, where D_j is the difference metric derived from the fitness function. This process was repeated for 30 generations. In summary, 30 × 2 × 1000 synthetic samples were compared with the sample via $P_j(x_j)$ comparisons for a given x_j to derive its respective h_j used in Eq. (2). A given X to Y map was then constructed with x_j sampled from Eq. (2) with h_j = E[terminal population of candidate h_j].

As a modification, the Kolmogorov Smirnov (KS) test[32] was used as the fitness function in the DE optimization. This is a nonparametric test that can be used to compare two numerically derived cumulative probability functions or compare a numerically derived curve to a reference[32]. Here we compare the respective numerical univariate cumulative probability functions derived from synthetic samples with those derived from the sample. The difference metric, D_j, for the KS test is the absolute maximum difference between the two cumulative probability functions under comparison.

Mapping. For each map in Step 1, we solve $P_j(x_j) = G_j(y_j)$ numerically with interpolation methods described previously[33,34], where $P_j(x_j)$ and $G_j(y_j)$ are assumed to be monotonically increasing. This solves the random variable transformation for each y_j analogous to histogram matching with double precision accuracy. Synthetic y_j ($n = 10^6$) were generated as standardized normal random variables using the Box-Muller (BM) method. Maps from X to Y are expressed as $y_j = m_j(x_j)$, where m_j is the jth map. The corresponding inverse maps, $x_j = m_j^{-1}(y_j)$, were derived numerically by inverting a given map and solving for x_j. The map construction was complemented by generating synthetic x_j with Eq. (2) using h_j derived from DE optimization. Synthetic x_j generated here were not used further.

Synthetic population generation. Synthetic populations (SPs) were generated in the uncorrelated T representation and converted back to X via Y. In Step 2, the PCA transform for the sample is given by

$$\mathbf{T} = \mathbf{YP}, \tag{3}$$

where $\mathbf{P}$ is a d × d matrix with uncorrelated normalized columns. These are the normalized eigenvectors of $\mathbf{C}_y$ that capture the sample's covariance structure. $\mathbf{C}_t$ is diagonal with: $c_{jj} = \sigma_j^2(t)$ corresponding to the ordered eigenvalues of $\mathbf{C}_y$. We make the approximation that $r(\mathbf{t})$ from the sample has the multivariate normal form expressed as

$$r(\mathbf{t}) = (2\pi)^{-\frac{d}{2}} |\mathbf{C}_t|^{-\frac{1}{2}} \exp - \frac{1}{2}\left[\mathbf{t}^T \mathbf{C}_t^{-1} \mathbf{t}\right]. \tag{4}$$

When the multivariate normality approximation holds in Y, it should hold in T. In Step 3, synthetic t_j were populated as zero mean independent normally distributed random variables with variances = $\sigma_j^2(t)$ using the BM method, producing the SP in T ($\mathbf{T}_s$). The row length of T_s defines the number of realizations in each SP and is

a

	Area (x_1)	Minor axis length (x_2)	Eccentricity (x_3)	Convex area (x_4)	Equivalent diameter (x_5)	Extent (x_6)	Solidity (x_7)	Roundness (x_8)	Shape factor 3 (x_9)	Shape factor 4 (x_{10})
Area (x_1)	**1.94E+07**	3.22E+04	5.28E+01	1.97E+07	5.19E+04	−1.46E+01	7.32E−01	−2.23E+01	−6.35E+01	−1.35E+00
	(1.79e+07, 2.21e+07)	(2.83e+04, 3.63e+04)	(4.45e+01, 6.85e+01)	(1.81e+07, 2.23e+07)	(4.79e+04, 5.88e+04)	(−2.49e+01, 3.36e+00)	(−1.64e−01, 1.76e+00)	(−3.73e+01, −1.93e+01)	(−8.26e+01, −5.40e+01)	(−2.35e+00, −4.87e−01)
Minor axis length (x_2)	0.8060	**8.22E+01**	−7.09E−02	3.25E+04	8.61E+01	7.35E−03	2.49E−03	3.75E−02	8.55E−02	−1.31E−03
		(7.42e+01, 9.12e+01)	(−9.35e−02, −4.78e−02)	(2.87e+04, 3.68e+04)	(7.59e+01, 9.72e+01)	(−1.95e−02, 4.00e−02)	(6.19e−04, 4.68e−03)	(1.57e−02, 5.04e−02)	(5.75e−02, 1.13e−01)	(−3.37e−03, 2.79e−04)
Eccentricity (x_3)	0.3710	−0.2420	**1.04E−03**	5.38E+01	1.41E−01	−2.09E−04	−5.90E−06	−4.60E−04	−1.24E−03	−1.17E−05
			(9.46e−04, 1.19e−03)	(4.44e+01, 6.82e+01)	(1.17e−01, 1.82e−01)	(−2.97e−04, −8.59e−05)	(−1.31e−05, 1.13e−06)	(−5.93e−04, −4.24e−04)	(−1.42e−03, −1.14e−03)	(−1.74e−05, −4.76e−06)
Convex area (x_4)	1.0000	0.8040	0.3740	**1.99E+07**	5.24E+04	−1.55E+01	3.40E−01	−2.43E+01	−6.48E+01	−1.62E+00
				(1.84e+07, 2.26e+07)	(4.85e+04, 5.94e+04)	(−2.67e+01, 1.89e+00)	(−4.52e−01, 1.48e+00)	(−3.93e+01, −2.07e+01)	(−8.27e+01, −5.41e+01)	(−2.55e+00, −6.21e−01)
Equivalent diameter (x_5)	0.9990	0.8070	0.3720	0.9990	**1.38E+02**	−3.97E−02	2.04E−03	−5.85E−02	−1.70E−01	−3.62E−03
					(1.28e+02, 1.57e+02)	(−6.87e−02, 6.85e−03)	(−3.43e−04, 4.86e−03)	(−9.90e−02, −5.13e−02)	(−2.19e−01, −1.42e−01)	(−6.25e−03, −1.23e−03)
Extent (x_6)	−0.0765	0.0187	−0.1500	−0.0801	−0.0777	**1.88E−03**	1.55E−05	2.82E−04	2.66E−04	5.62E−06
						(1.75e−03, 2.03e−03)	(5.88e−06, 2.61e−05)	(1.78e−04, 3.54e−04)	(1.12e−04, 3.68e−04)	(−6.97e−06, 1.10e−05)
Solidity (x_7)	0.0565	0.0934	−0.0623	0.0260	0.0589	0.1220	**8.62E−06**	3.92E−05	1.17E−05	5.32E−06
							(6.71e−06, 1.04e−05)	(2.42e−05, 4.13e−05)	(1.57e−06, 1.90e−05)	(2.90e−06, 5.03e−06)
Roundness (x_8)	−0.1990	0.1630	−0.5610	−0.2140	−0.1960	0.2560	0.5250	**6.46E−04**	5.77E−04	2.10E−05
								(4.70e−04, 7.81e−04)	(5.28e−04, 7.41e−04)	(1.25e−05, 2.56e−05)
Shaped factor 3 (x_9)	−0.3720	0.2440	−0.9960	−0.3760	−0.3730	0.1590	0.1030	0.5870	**1.50E−03**	1.96E−05
									(1.38e−03, 1.70e−03)	(9.74e−06, 2.54e−05)
Shape factor 4 (x_{10})	−0.1130	−0.0533	−0.1350	−0.1340	−0.1140	0.0480	0.6710	0.3060	0.1880	**7.29E−06**
										(5.41e−06, 8.93e−06)

b

	Area (y_1)	Minor axis length (y_2)	Eccentricity (y_3)	Convex area (y_4)	Equivalent diameter (y_5)	Extent (y_6)	Solidity (y_7)	Roundness (y_8)	Shape factor 3 (y_9)	Shape factor 4 (y_{10})
Area (y_1)	**1.00E+00**	7.94E−01	3.86E−01	9.98E−01	9.99E−01	−5.43E−02	6.56E−02	−2.66E−01	−3.87E−01	−1.30E−01
	(8.98e−01, 1.11e+00)	(6.95e−01, 8.98e−01)	(3.06e−01, 4.71e−01)	(8.96e−01, 1.11e+00)	(8.97e−01, 1.11e+00)	(−1.32e−01, 1.73e−02)	(−1.04e−02, 1.42e−01)	(−3.50e−01, −1.89e−01)	(−4.71e−01, −3.08e−01)	(−2.09e−01, −4.61e−02)
Minor axis length (y_2)	0.7940	**1.00E+00**	−2.38E−01	7.97E−01	7.98E−01	2.91E−02	1.05E−01	1.51E−01	2.37E−01	−6.76E−02
		(8.94e−01, 1.10e+00)	(−3.15e−01, −1.62e−01)	(6.98e−01, 8.99e−01)	(6.98e−01, 9.00e−01)	(−4.55e−02, 1.03e−01)	(2.92e−02, 1.79e−01)	(7.92e−02, 2.27e−01)	(1.61e−01, 3.15e−01)	(−1.44e−01, 1.37e−02)
Eccentricity (y_3)	0.3860	−0.2380	**1.00E+00**	3.81E−01	3.81E−01	−1.39E−01	−7.12E−02	−6.64E−01	−9.98E−01	−1.40E−01
			(8.92e−01, 1.11e+00)	(3.03e−01, 4.66e−01)	(3.02e−01, 4.66e−01)	(−2.17e−01, −6.71e−02)	(−1.46e−01, 7.79e−03)	(−7.61e−01, −5.72e−01)	(−1.11e+00, −8.92e−01)	(−2.15e−01, −6.29e−02)
Convex area (y_4)	0.9980	0.7970	0.3810	**1.00E+00**	9.99E−01	−6.17E−02	4.14E−02	−2.80E−01	−3.84E−01	−1.44E−01
				(8.96e−01, 1.11e+00)	(8.96e−01, 1.11e+00)	(−1.40e−01, 9.80e−03)	(−3.58e−02, 1.19e−01)	(−3.62e−01, −2.02e−01)	(−4.68e−01, −3.05e−01)	(−2.24e−01, −5.87e−02)
Equivalent diameter (y_5)	0.9990	0.7980	0.3810	0.9990	**1.00E+00**	−5.74E−02	6.85E−02	−2.65E−01	−3.82E−01	−1.27E−01
					(8.97e−01, 1.11e+00)	(−1.36e−01, 1.46e−02)	(−8.13e−03, 1.47e−01)	(−3.48e−01, −1.86e−01)	(−4.66e−01, −3.03e−01)	(−2.07e−01, −4.24e−02)
Extent (y_6)	−0.0543	0.0291	−0.1390	−0.0617	−0.0574	**1.00E+00**	1.35E−01	2.62E−01	1.43E−01	1.40E−02
						(8.90e−01, 1.11e+00)	(6.08e−02, 2.15e−01)	(1.88e−01, 3.42e−01)	(7.14e−02, 2.23e−01)	(−5.91e−02, 9.98e−02)
Solidity (y_7)	0.0656	0.1050	−0.0712	0.0414	0.0685	0.1350	**1.00E+00**	4.83E−01	9.87E−02	5.42E−01
							(8.95e−01, 1.11e+00)	(3.99e−01, 5.75e−01)	(2.02e−02, 1.73e−01)	(4.52e−01, 6.35e−01)
Roundness (y_8)	−0.2660	0.1510	−0.6640	−0.2800	−0.2650	0.2620	0.4830	**1.00E+00**	6.79E−01	3.13E−01
								(8.94e−01, 1.12e+00)	(5.87e−01, 7.77e−01)	(2.33e−01, 3.97e−01)

Continued

b										
	Area (y_1)	**Minor axis length (y_2)**	**Eccentricity (y_3)**	**Convex area (y_4)**	**Equivalent diameter (y_5)**	**Extent (y_6)**	**Solidity (y_7)**	**Roundness (y_8)**	**Shape factor 3 (y_9)**	**Shape factor 4 (y_{10})**
Shaped factor 3 (y_9)	−0.3870	0.2370	−0.9980	−0.3840	−0.3820	0.1430	0.0987	0.6790	**1.00E+00**	1.82E−01
									(8.94e−01, 1.11e+00)	(1.04e−01, 2.58e−01)
Shape factor 4 (y_{10})	−0.1300	−0.0676	−0.1400	−0.1440	−0.1270	0.0140	0.5420	0.3130	0.1820	**1.00E+00**
										(8.92e−01, 1.12e+00)

Table 3. Covariance and correlation for Dataset 3: in both tables, entries on the diagonals and above give covariance quantities. Entries below the diagonals (bold) provide the respective Pearson correlation coefficients. Table 3a gives the X representation quantities and 3b the Y representation quantities. The covariance quantities were generated from the respective sample. Parenthetically, 95% confidence intervals generated from synthetic samples are cited below the respective covariance quantity. Variables are cited with the names used in their respective resource and with the names used in this report parenthetically.

arbitrary. Here, we let $n = 10^6$ for all SPs. In Step 4 to construct the SP in Y ($\mathbf{Y}_s$), the inverse PCA transform was used by substituting $\mathbf{T}_s$ for $\mathbf{T}$ in Eq. (3) giving

$$\mathbf{Y}_s = \mathbf{T}_s\mathbf{P}^T. \tag{5}$$

With this process, $g(\mathbf{y}) = g_n(\mathbf{y})$ for synthetic data. By premise, the covariance of $\mathbf{Y}_s$ should be like that of $\mathbf{Y}$. In Step 5 to produce the SP in X ($\mathbf{X}_s$), synthetic y_j were inverse mapped. Similarly, the covariance of $\mathbf{X}$ should be like that of $\mathbf{X}_s$. An example is also provided to illustrate that $\mathbf{X}_s$ is densely populated in contrast with the sparse sample in X.

Statistical methods. The goals are to evaluate the latent normal characteristic and to produce synthetic data that is statistically like its sample. This analysis is based on both multiple univariate/multivariate pdf and covariance comparisons. A given synthetic sample (n = 667) was drawn at random from its SP. The same realizations from a given synthetic sample were used for comparisons in X, Y, and T, when applicable.

Probability density function comparisons. Univariate pdf comparisons. The KS test (described in Step 1) was used for all univariate pdf comparisons. For such comparisons, we selected the test threshold at the 5% significance level as the critical value. In X, $p_j(x_j)$ from the sample were tested for normality and compared with their respective pdfs from synthetic samples created in Step 5. In Y, $g_j(y_j)$ from the sample were compared with the respective pdfs from their synthetic samples produced by Step 4; this implicitly evaluated univariate normality in Y because synthetic y_j were derived from a multivariate normal process in T_s. In T, we compared $r_j(t_j)$ from the sample with their respective pdfs derived from zero mean normally distributed random variables (i.e., synthetic t_j) with variances $= \sigma_j^2(t)$ [from Step 3]. For each t_j, the sample was compared with 1000 synthetic samples, and the percentage of times that measured D_j was less than the critical test value was tabulated.

Distribution free multivariate pdf comparisons. To evaluate whether the sample and synthetic samples were drawn from the same distribution without assumptions, we used the maximum mean discrepancy (MMD)[35] test. This is a kernel-based (normal kernel) analysis that computes the difference between every possible vector combination between and within two samples (excluding same vector comparisons). To determine the kernel parameter for these tests, we used the median heuristic[35,36]. This analysis is based on the critical value (MMD_c) at the 5% significance level and the test statistic (MMD_u^2). Both quantities are calculated from the two samples under comparison. This test has an acceptance region given the two distributions are the same: $MMD_u^2 < MMD_c$ (see Theorem 10 and Corollary in[35]). This test was applied in X, Y, and T. Note, when applying this test in either T or Y, it is implicitly testing the sample's likeness with multivariate normality. In X, Y, and T, 1000 synthetic samples were compared with the sample. The test acceptance percentage was tabulated. MMD_u^2 and MMD_c values are provided as averages over all trials because they change per comparison.

Random projection multivariate normality evaluation. Random projections were used to develop a test for normality. The vector $\mathbf{w}$ with d components is multivariate normal if the scalar random variable, $z = \mathbf{u}^T\mathbf{w}$, is univariate normal, where $\mathbf{u}$ is a d component vector with unit norm that is defined as a projection vector in this report[37,38]. As mentioned by Zhou and Saho[38], we developed this formulism into a specific random projection test. To actualize such a test to probe the samples and synthetic samples similarity with normality, the projection vector $\mathbf{u}$ was generated randomly 1000 times, referenced as $\mathbf{u}_s$. Here, s is the projection index ranging from [1,1000]. The projection equation is then expressed as

$$z|s = \mathbf{u}_s^T\mathbf{w}, \tag{6}$$

where $z|s$ defines the scalar z conditioned upon s. In Eq. (6), $\mathbf{x}$, $\mathbf{y}$, or $\mathbf{t}$ was substituted for $\mathbf{w}$, and given projection was taken over all realizations (i.e., n = 667) of given sample. These realizations of $z|s$ were used to form the normalized histogram that approximates the conditional pdf for the left side of Eq. (6) defined as $f(z|s)$. A

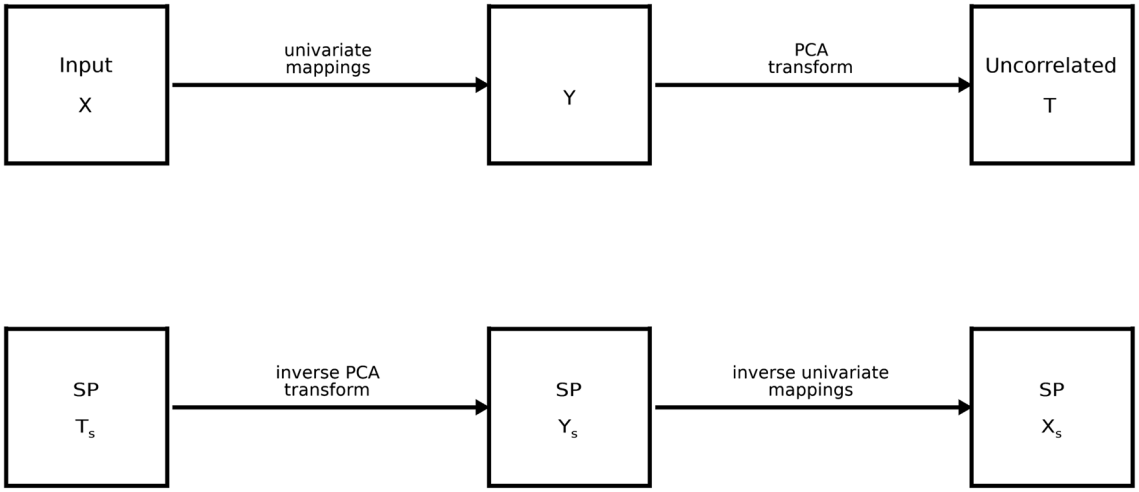

Figure 4. Processing Flow: the top row shows the processing flow for the sample. The reversed processing flow for the synthetic population generation is shown in the bottom row.

different series of $\mathbf{u}_s$ was produced for each representation; once a given series was produced, $\mathbf{u}_s$ remains fixed. The components of $\mathbf{u}_s$ were generated as standardized normal random variables, where $\mathbf{u}_s$ was normalized to unit norm. For a given sample, f(z|s) was tested for normality using the KS test. This procedure was repeated for all random projections (all s), resulting in 1000 KS test comparisons for normality. The percentage of the times that the null hypothesis was not rejected was tabulated as the normality similarity gauge. We refer to this procedure as the *random projection test*. It is the percentage of times that **w** was not rejected when probed in 1000 random directions. This test was performed once for each sample and with 100 synthetic samples and averaged. Here, the same $\mathbf{u}_s$ series used to probe a given sample was also used to probe 100 of its respective synthetic samples. We note, synthetic samples are multivariate normal in Y and T by their construction. Tests were performed on synthetic samples (Y and T) to give control standards as normal comparators. Tests were also performed in X: (1) as control comparator for the test itself; and (2) to determine if a given sample was multivariate normal before undergoing the mapping.

To gain both insight into Eq. (6) and the test, expressions for both z|s and f(z|s) are developed. First, z|s results from a linear random variable transform given by

$$z|s = u_1 w_1 + u_2 w_2 + u_k w_k + \cdots + u_d w_d, \tag{7}$$

where u_k are the components of $\mathbf{u}_s$, and $h_k(w_k)$ are the univariate pdfs for scaled w_k. To check one endpoint, we assume w_k are independent as a coarse approximation to our samples. Then, f(z|s) results from repeated convolutions given by

$$f(z|s) = c \times ((((h_1 * h_2) * h_3) * h_4) * \ldots) * h_d, \tag{8}$$

where $c^{-1} = u_1 \times u_2 \times u_3 \times u_4 \times \ldots \times u_d$, and $h_k(w_k) \sim h_k$. If some $h_k(w_k)$ have relatively much larger variances (widths) than others, their functional forms can tend to predominate Eq. (8).

Mardia multivariate normality test. This is a two component test that uses multivariate skewness and kurtosis for evaluating deviations from normality[37], applied in X, Y and T. It produces a deviation measure for each component as well as a combined measure; we cite the component-findings. This test was applied in X as a control. Outlier elimination techniques were not applied.

Covariance comparisons. Two methods were used to evaluate the covariance similarity between samples and synthetic samples: with (CIs) and eigenvalue comparisons.

Comparisons with confidence intervals. Each covariance matrix element between the sample and its respective synthetic samples was compared with CIs. We assumed the sample and synthetic samples were drawn from the same distributions. We used the elements from each $\mathbf{C}_x$ and $\mathbf{C}_y$ as point estimates from a given sample in both X and Y. One thousand synthetic samples (n = 667) were used to calculate 1000 covariance matrices (in X and Y). For each matrix element, the respective univariate pdf was formed, and 95% CIs were calculated. This procedure was repeated 1000 times. The percentage of times the sample's point estimate (for each element in $\mathbf{C}_y$ and $\mathbf{C}_x$) was within the synthetic element's CIs was tabulated.

Comparisons with PCA. The eigenvalues from $\mathbf{C}_y$ were used as the reference comparators under two conditions. For condition 1, the PCA transform determined with the sample was applied to a synthetic sample (sample/syn test). Synthetic eigenvalues were estimated by calculating the variances of synthetic t_j. For condition

DS1			DS2			DS3		
Variable name	Mean	Standard deviation	Variable name	Mean	Standard deviation	Variable name	Mean	Standard deviation
P_1 (x_1)	1.48E+01	1.51E+01	Mean (x_1)	5.35E+02	1.24E+02	Area (x_1)	4.48E+04	4.41E+03
P_2 (x_2)	5.83E+00	5.27E+00	SD (x_2)	6.33E+01	4.37E+01	Minor Axis Length (x_2)	1.91E+02	9.06E+00
P_3 (x_3)	3.23E+00	2.80E+00	HP Filter (x_3)	4.01E−02	1.32E−02	Eccentricity (x_3)	7.69E−01	3.23E−02
P_4 (x_4)	2.00E+00	1.59E+00	LP Filter (x_4)	6.74E+00	1.64E+00	Convex Area (x_4)	4.53E+04	4.46E+03
P_5 (x_5)	1.32E+00	9.79E−01	Local SD (x_5)	4.67E+01	2.78E+01	Equivalent Diameter (x_5)	2.39E+02	1.18E+01
P_6 (x_6)	9.21E−01	6.74E−01	P_{20} (x_6)	1.20E−01	7.39E−02	Extent (x_6)	7.58E−01	4.33E−02
P_7 (x_7)	6.72E−01	4.71E−01	Local Correlation (x_7)	3.02E−02	5.67E−03	Solidity (x_7)	9.88E−01	2.94E−03
P_8 (x_8)	5.16E−01	3.54E−01	Br Area (x_8)	1.77E+02	7.84E+01	Roundness (x_8)	8.84E−01	2.54E−02
Age (x_9)	5.83E+01	1.17E+01	Age (x_9)	5.83E+01	1.17E+01	Shaped Factor 3 (x_9)	6.35E−01	3.87E−02
BMI (x_{10})	2.79E+01	6.73E+00	BMI (x_{10})	2.79E+01	6.73E+00	Shaped Factor 4 (x_{10})	9.95E−01	2.70E−03

Table 4. Mean and standard deviations: this gives the univariate distribution means and standard deviations for all variables by dataset in the X representation. Variables are cited with the names used in their respective resource and with the names used in this report parenthetically.

2, the PCA transform determined with a synthetic sample selected at random was applied to the sample (syn/sample test). Eigenvalues were estimated by calculating the variances of t_j from the sample. For both conditions, each eigenvalue (or equivalently, variance) was compared to its respective reference (sample) using the F-test.

Ethics and consent to participate. All methods were carried out in accordance with relevant guidelines and regulations. All experimental procedures were approved by the Institutional Review Board (IRB) of the University of South Florida, Tampa, FL under protocol #Ame13_104715. Mammography data was collected retrospectively on a waiver for informed consent approved by the IRB of the University of South Florida, Tampa, FL under protocol #Ame13_104715.

Results

Univariate normality analysis in the X representation. Figures 1, 2 and 3 show the univariate pdfs (solid) for each sample in X. Of note, many pdfs are observably non-normal, usually right skewed. Each x_j in DS1 showed significant deviation from normality ($p < 0.0001$) except for x_9. In DS2, neither x_1 or x_9 (x_9 is the same in DS1) showed significant deviation from normality, while the remaining x_j exhibited significant deviations ($p < 0.0003$) except for x_3 ($p = 0.0144$). In DS3, x_1 through x_5 and x_9 did not show significant deviations from normality, the remaining x_j deviated significantly ($p < 0.002$).

Mapping and KDE optimization. *Mapping.* Figure 5 shows an example of the X to Y map for y_9 in the left-pane and the inverse Y to X map in the right-pane. Red-dashed lines show the map and its inverse constructed without synthetic x_j. Staircasing effects are observable particularly in the tail regions, where sample densities are sparse. Black lines show the map and its inverse constructed with $n = 667$ (the sample) plus $n = 10^6$ synthetic realizations produced with optimized uKDE. Staircasing effects were removed when incorporating synthetic x_j, which was common with all maps and inverses (not shown).

uKDE optimization. The optimization produced bandwidth parameter solutions (h_j) used in Eq. (2). Here, we illustrate the evolution of the *solution* with h_9 from DS1 and DS2 as an example. Figure 6 shows the scatter plot between the candidate h_9 population and the respective D_9 (KS test difference metric) for DE generation = 1 in the left-pane and for the terminal generation = 30 in the middle-pane. The solution space (middle-pane) is tightly clustered indicating DE *convergence*. A closer view of this cluster is shown in the right-pane of Fig. 6. This relatively tight-cluster characteristic was common among all variables and datasets (not shown).

Univariate comparisons between samples and synthetic samples. *Comparisons in T.* These findings are summarized first because they start the flow back to X and can show departures from normality. Figures 7, 8 and 9 show the pdfs for the samples (solid) compared with their corresponding synthetic pdfs (dashed). Table 5 shows the variances in T for each sample (i.e., eigenvalues for each sample). Due to (1) the normalization in Y, and (2) that $d = 10$, multiplying a given, $\sigma_j^2(t)$ by 10% gives the percentage of the total variance explained by its t_j. Table 6 shows the KS test findings for the univariate normality comparisons. Here we use a cutoff of < 65% to indicate deviation as most trends were well above this boundary. As shown in Table 6 (left column for each dataset): (1) the normal model did not deviate for any t_j in DS1 (7 t_j were < 94%); (2) the normal model deviated for t_{10} in DS2 (5 t_j were < 94%); and (3) the normal model deviated for t_7, t_8, t_9, and t_{10} in DS3 (3 t_j were < 94%). In DS2, t_{10} explains about 0.2% of the total variance. Similarly, in DS3, the sum of the variances of the four variables (t_7, t_8, t_9, and t_{10}) constituted about 0.14% of the total variance.

Comparisons in Y. Figures 10, 11 and 12 show the pdfs in Y resulting from the mapped samples (i.e., mapped x_j) for each dataset (solid) compared with their respective synthetic pdfs (dashed), which are normal by construct. Comparisons in Y showed little departure from normality in any sample, as the tests were not rejected (about 99%) in most instances (Table 6, middle column for each dataset). Findings from DS2 and DS3 indicate that substituting normal pdfs in T whenever sample t_j deviated from normality had little influence on this analysis. This may be because these respective variables in total or isolation explained a minute portion of the variance in the respective PCA models.

Comparisons in X. Figures 1, 2 and 3 show the pdfs in X for the samples (solid) compared with their corresponding synthetic pdfs (dashed). The pdfs from the sample did not deviate from their corresponding synthetic pdfs in any dataset, as the tests were not rejected (< 99%) in most instances (Table 6, right columns). The parenthetical entries in Table 6 show the test findings without using synthetic data for the map/inverse constructions (using the samples only). These show that complementing the map constructions with uKDE is a necessary component of this methodology, although the degree of deviation from the KS tests varied across datasets. Note, the improvement held in D3 as well, which was normal in X (shown below).

Multivariate comparisons and normality comparisons. *MMD tests.* Testing was performed in X, Y, and T, and the test metrics are provided in Table 7. These show the samples and respective synthetic samples were drawn from the same distributions. In 100% the tests, measured MMD_u^2 values were less than the critical MMD_c quantities. The MMD tests in Y and T were also proxy tests for sample-normality due to the SP constructs.

Random projection normality tests. Testing was performed in X, Y, and T, and the findings are shown in Table 8. Test findings were mixed for the samples in X and were not rejected for approximately these instances: 40% in DS1; 74% in DS2; and 99% in DS3. Thus, DS3 is better approximated as normal in X compared to the other samples. The tests for synthetic samples in X tracked the findings for their respective samples: 44%, 74% and 99% respectively. In Y, the tests for the samples were not rejected for about these instances: 99% in DS1, 97% in DS2, and 95% in DS3, whereas the test for the synthetic samples should no deviation from normality. Similarly in T, the tests for the samples were not rejected for about these instances: 99% in DS1, 94% in DS2, and 95% in DS3. There is a difference in the X and Y analyses because the mapping normalizes the variables in Y. In DS3, the standard deviations vary over many orders of magnitude (see Table 4). As shown by Eq. (8), variables in DS3 with the larger standard deviations may wash out the other variables; the variables in X that had normal marginals compared with those that were not, indicates that a portion of the normal marginals had much larger standard deviations (in Table 4, see x_1–x_4). As another control experiment, we standardized all variables in X to zero mean and unit variance and performed the tests again. The tests for the samples gave: 32.9%, 76.4%, and 75.2% for DS1, DS2, and DS3, respectively. For synthetic samples, these tests gave: 33.6%, 80.1% and 89.1% for DS1, DS2, and DS3, respectively. Note, centering the means alone had no influence on the findings as expected (data not shown). Thus, normalizing the univariate measures can influence the likeliness with normality by virtue of Eq. (8). In sum, these tests show all samples resemble multivariate normality in both Y and T and that the sample for DS3 resembles normality in X without mapping (without first normalizing the variances).

Mardia normality tests. Testing was performed in X, Y, and T. The findings are shown in Table 8. In X, the samples and synthetic samples all deviated from normality (both skewness and kurtosis). In both Y and T, the samples showed significant deviations from normally in all tests. In contrast, synthetic samples did not deviate significantly from normality in any test in Y or T, as expected.

Covariance comparisons. *Covariance matrix comparisons with confidence intervals.* Test findings are provided in Tables 1, 2 and 3 for the respective datasets. Part-a of each table shows the X quantities, and part-b shows the corresponding Y quantities. For DS1 (Table 1), covariance references (sample) were within the CIs of the synthetic data for 100% of the trials in both X and Y. For DS2 (Table 2), most references agreed with the synthetic elements except for two entries in X (Table 2a). From the 1000 trials, the x_2x_3 covariance was out of tolerance for 0.1% of the instances, and the x_5x_{10} covariance was out for 18.7% of instances. For DS3, all covariance references were within tolerance except the x_7x_{10} covariance, which was out of tolerance for 100% of the instances. In tests that showed more deviation (percentage > 0.1%), the reference covariances were approximately zero.

Eigenvalue comparison tests. Eigenvalues are provided in Table 5. This table is separated into three sections vertically. Reference eigenvalues are provided in the top row of each section. Eigenvalues calculated from the sample/syn (condition 1) and syn/sample (condition 2) are provided in the middle and bottom rows of each section, respectively. F-tests were not significant ($p > 0.05$) in any comparison with the references indicating similarity.

Sample sparsity and synthetic population space filling. This illustration demonstrates that the approach fills in the multidimensional space with synthetic realizations derived from a relatively sparse sample. We selected a synthetic entity at random from DS1 giving this vector: $\mathbf{x}^T = [4.20, 2.08, 1.61, 1.15, 0.85, 0.67, 0.54, 0.44, 52.0, 23.6]$. We selected x_1 and x_8 as the scatter plot variables. For the other 8 components, all synthetic realizations within $x_{ij} \pm \frac{1}{2}\,\sigma_j$ (the standard deviation for x_j) were selected and viewed in the x_1x_8 plane as a scatter plot. The same vector and limits were used to select realizations from the sample. The plots are provided in Fig. 13 for comparison. The sample (left-pane) produced n = 24 realizations, whereas the SP (right-pane) pro-

duced n = 36,398 realizations. These plots illustrate the sample's relative sparsity and that the synthetic approach produces a dense population with observations that did not exist in the sample.

Discussion

The work involved several steps to generate synthetic data from arbitrarily distributed samples. To the best of our knowledge, new aspects and findings from this work include: (1) demonstrating a class of arbitrarily distributed samples has a latent normal characteristic, as exhibited by two of the samples; (2) conditioning the input variables with sample size augmentation and then constructing univariate transforms so that known techniques could be applied to generate synthetic data; (3) deploying multiple statistical tests for assessing both normality and general similarity in both the univariate and multivariate pdf settings; (4) developing methods for comparing covariance matrices; and (5) incorporating differential evolution (DE) optimization for uKDE bandwidth determination based on the KS fitness function. The related findings are discussed below in detail.

A method was presented that converts a given multivariate sample into multiple 1D marginal pdfs by constructing maps. These X–Y maps were constructed by augmenting the sample size with optimized uKDE. Performing the analysis with and without data augmentation improved the marginal pdf comparisons between the samples and synthetic samples; this also held in DS3 (four x_j), which was approximately normal in X. PCA applied to standardized normal variables in Y produced uncorrelated variables in T, where synthetic data was generated. This approach essentially decouples the problem into the covariance relationships (in **P** and its inverse) and 1D marginal pdfs (i.e., approximate parametric models in Y and T). This decoupling is similar to the objective of Copula modeling that follows from Sklar's work[39,40]. Copula modeling allows specification of marginal pdfs and the correlation structure independently[41]; in this approach, the marginals must be specified accurately and finding analytical solutions for $d > 4$ is difficult[42]. In contrast with Copula modeling, which is flexible, the covariance (or correlation) structure in our approach is fixed by the normal form and empirically derived; the marginals were forced to normality rather than specified. As a benefit, the CIs for $\mathbf{C}_k$ with our approach were estimated from the pdfs for each matrix element without assumption other than the normal calculation form. Additionally, the eigenvalue comparison technique results reinforced the CI comparison findings. Outside of the multivariate normal situation it is not clear when (1) comparing the marginals with one set of tests, and (2) comparing the covariance relationships separately with another set of tests results in a good overall empirical comparison-approximation between two multivariate samples. Such situations will require further analyses.

There are several other points worth noting about this work. An empirically driven stochastic optimization technique was used to estimate the uKDE bandwidth parameters for the map/inverse constructions. The relative efficiency of the approach is an important attribute in that it only requires multiple uKDE applications rather than mKDE. The number of generations in the optimization was fixed. This can be changed easily to a variable termination based on achieving a critical threshold or applying other appropriate fitness functions; for example, the stopping criteria could be based on the critical distance in the KS test or the change in this distance from one generation to the next. Likewise, there are plug-in kernel bandwidth parameters that can be used statically. These are derived by considering closed form expressions containing the constituent pdfs and minimizing the asymptotic behavior of either the mean integrated square error or mean squared error[43]. We explored such parameters[44], but they did not perform as well as the KS test with DE, notwithstanding the number of computations used here to determine a given bandwidth parameter. Of note, the KS test has limitations, as it is more sensitive to the median of the distribution rather than the tails. As an alternative, the Anderson Darling test is a variant of the KS procedure that is sensitive to the distribution tails[32]. The mapping from X to Y standardized the problem at the univariate level, but in general there is no guarantee that collectively it produced a multivariate normal in Y. Testing performed in Y (Step 2) could be used to discriminate input samples that have the latent normal characteristic from those that do not. The random project test could be developed into a gauge at this step for assessing the deviation from normality. Moreover, comparing $r(t_j)$ from the sample against normality (following Step 3) also provides a basis for testing sample's likeness to a multivariate normal (discussed below). When $p(\mathbf{x})$ is approximately multivariate normal, as in DS3, the mapping is not required and generating synthetic data based on PCA (without the mapping step) is a practiced technique; our approach addresses the case when this approximation fails to hold. The random projection tests changed the similarity with normality when standardizing the samples. Thus, the purpose for generating synthetic data should be considered before adjusting the input sample.

There are several other limitations and qualifications worth noting. Several multivariate pdf tests were examined with mixed findings. MMD tests in X, Y, and T showed each sample was statistically similar with its respective synthetic sample(s). These tests also indicated normality in Y and T (by default). This MMD test is sensitive to changes in the mean. In our processing, all means were forced either to identically zero or to statistical similarity via mapping. Likewise, the heuristic used for the kernel bandwidth determination can be less than optimal under certain conditions, decreasing the MMD test performance[45]. Random projection tests in Y and T indicated that the samples did not deviate from normality in most instances, whereas synthetic samples showed essentially no deviation. Understanding the acceptable departure from normality for this test in the modeling context will require more work. This test also showed DS3 was approximately normal in X. In contrast, Mardia tests showed all samples deviated significantly from normality in X, Y and T. With the Mardia test, synthetic samples showed: (1) essentially no deviation in Y or T as expected; (2) and significant deviation in X. Here, we made no attempt to mitigate possible outlier interference when analyzing the samples[46]. Note, testing for multivariate normality is not a trivial task; many of the complexities are covered by Farrell et al.[47]

The conclusions we make from these tests indicated each sample was approximately multivariate normal in Y and T, noting the approach may not be dependent upon this characteristic as elaborated below. In planned research, these approximations will be tested in the modeling context to evaluate whether sample and synthetic

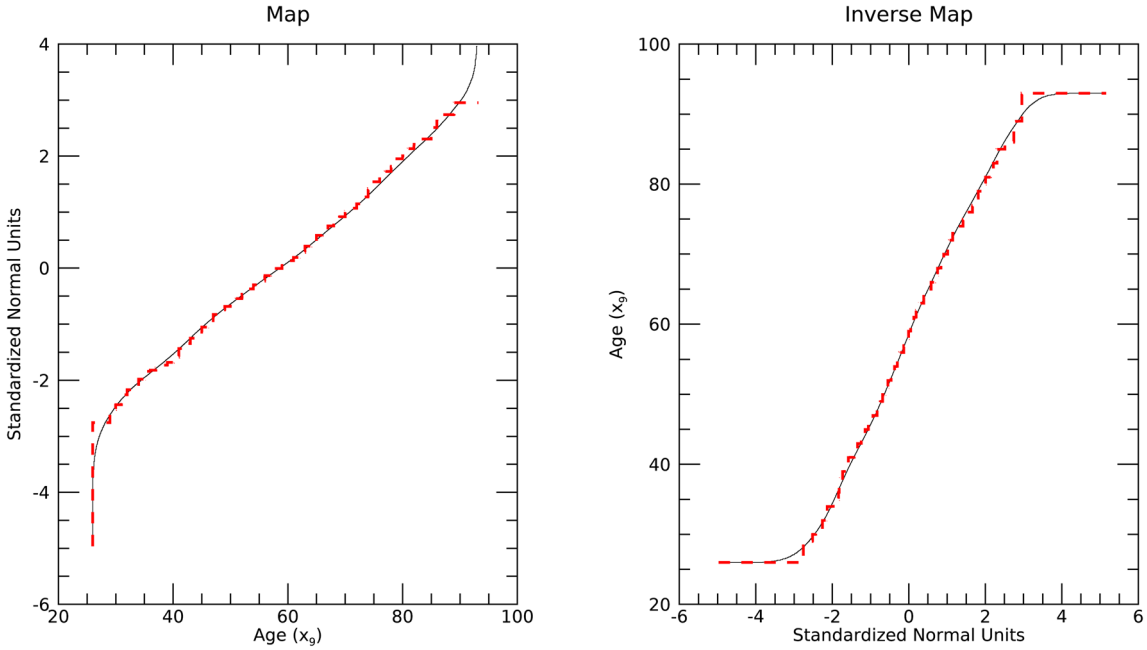

Figure 5. Univariate Mapping Illustration: this shows the map (left) and inverse map (right) for age (x_9) used in both DS1 and DS2. Maps using the sample only (red-dashes) are compared with maps augmented with synthetic data (black-sold).

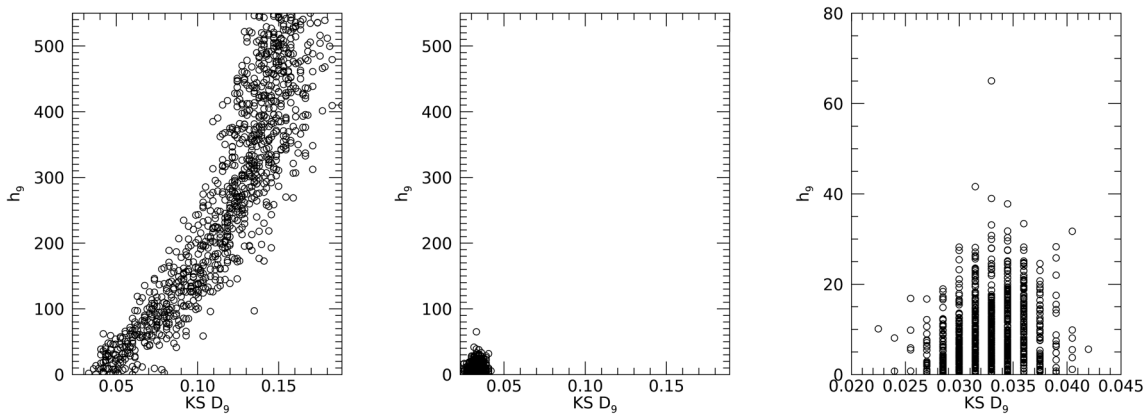

Figure 6. Kernel Density Estimation Optimization Illustration: this shows the differential evolution optimization to determine h_9 (for x_9, age from DS1 and DS2). These show the scatter plots of the Kolmogorov Smirnov test metric (measured D_9 quantities for entire generation) versus the h_9 quantities for two generations: generation = 1 (left); terminal generation = 30 (middle); and closeup view of the terminal generation (right). The terminal generation shows the candidate solutions for h_9 are tightly clustered (compare left pane with middle and right panes).

data are interchangeable. When this normality approximation holds, it implies that the original multivariate pdf estimation problem in X was converted to a parametric normal model described by Eq. (1), which simplifies the synthetic data generation. If this conversion generalizes to other datasets (at least in part), it implies that some class (subset) of the multivariate sample space can be studied with simulations by altering, n, d, and the covariance matrix to that of an arbitrary sample. Future work involves investigating arbitrary selected samples to understand how often this latent normal characteristic is present.

Alternatively, analyzing the samples in T may provide another method for comparing datasets, evaluating similarity, evaluating normality in Y, or generalizing our approach. The marginals from each sample were approximated as univariate normal in T, although there were noted variations. For example, and as noted, about 99% of the total variance came from the first four variables in DS1 (see Table 5). DS2 and DS3 were found to be similar with the first four variables accounting for 90–92% of the total variance. Thus, DS1 is more compressible

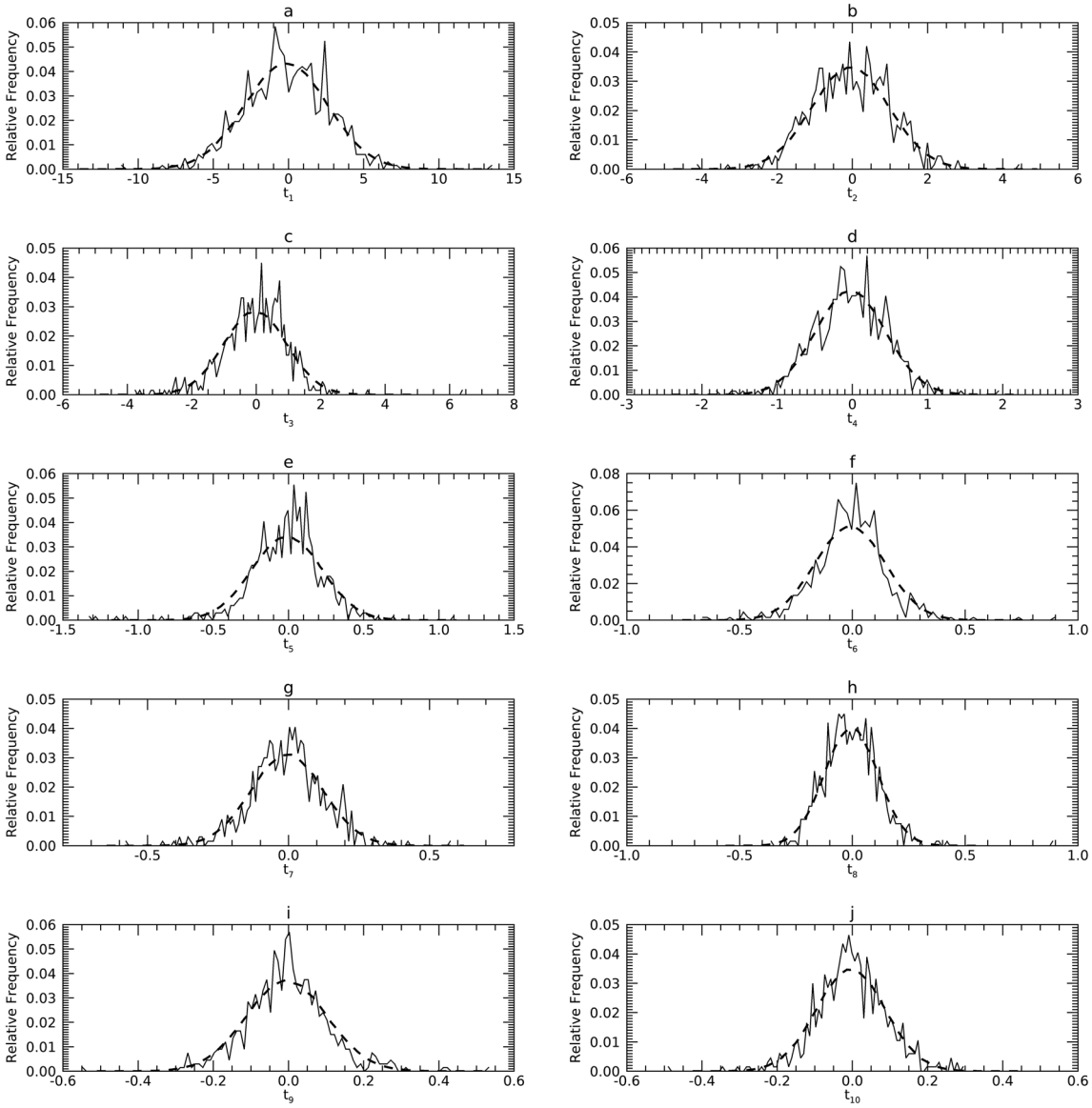

Figure 7. Marginal probability density functions (pdfs) for DS1 in the T representation: each pdf for DS1 (sold) is compared with its corresponding pdf from synthetic data (dashes).

than the other two datasets as expected due to the high correlation from its approximate functional Fourier form. Although not the purpose of this report, the amount of compression is a likely metric for estimating the effective dimensionality (d_e) when $d_e \leq d$, which could be useful for estimating sample size. When viewing the PCA transform through the NIPALS algorithm[14], it is clear when the total variance is explained by a number of components $d_e << d$, the remaining components are residue (noise, chatter, rounding errors). This effect could explain why deviations from normality in DS2 and DS3 in T did not influence the multivariate normal approximation in Y. Here, we did not encounter non-normal variables (from the samples) in T that explained a significant portion of the total variance. When a given sample is well approximated as multivariate normal in Y, the PCA transformation will produce univariate normal marginal pdfs in T. This step could be developed into the *definitive* test for multivariate normality in Y by understanding the residual error of the non-normal marginal pdfs in T. In this work, the analysis in T supports the normality findings for each sample because the residual non-normal errors were parasitic. Future work will investigate: (1) the impact of the residual error in T on normality in Y, and (2) causes for normality in T, i.e., possibly due to forced normality in Y, some characteristic of the X representation data, or the PCA transform. If required, the technique could be generalized to accommodate non-normal marginals in T. As a generalization, uKDE will be investigated for generating univariate non-normal distributions

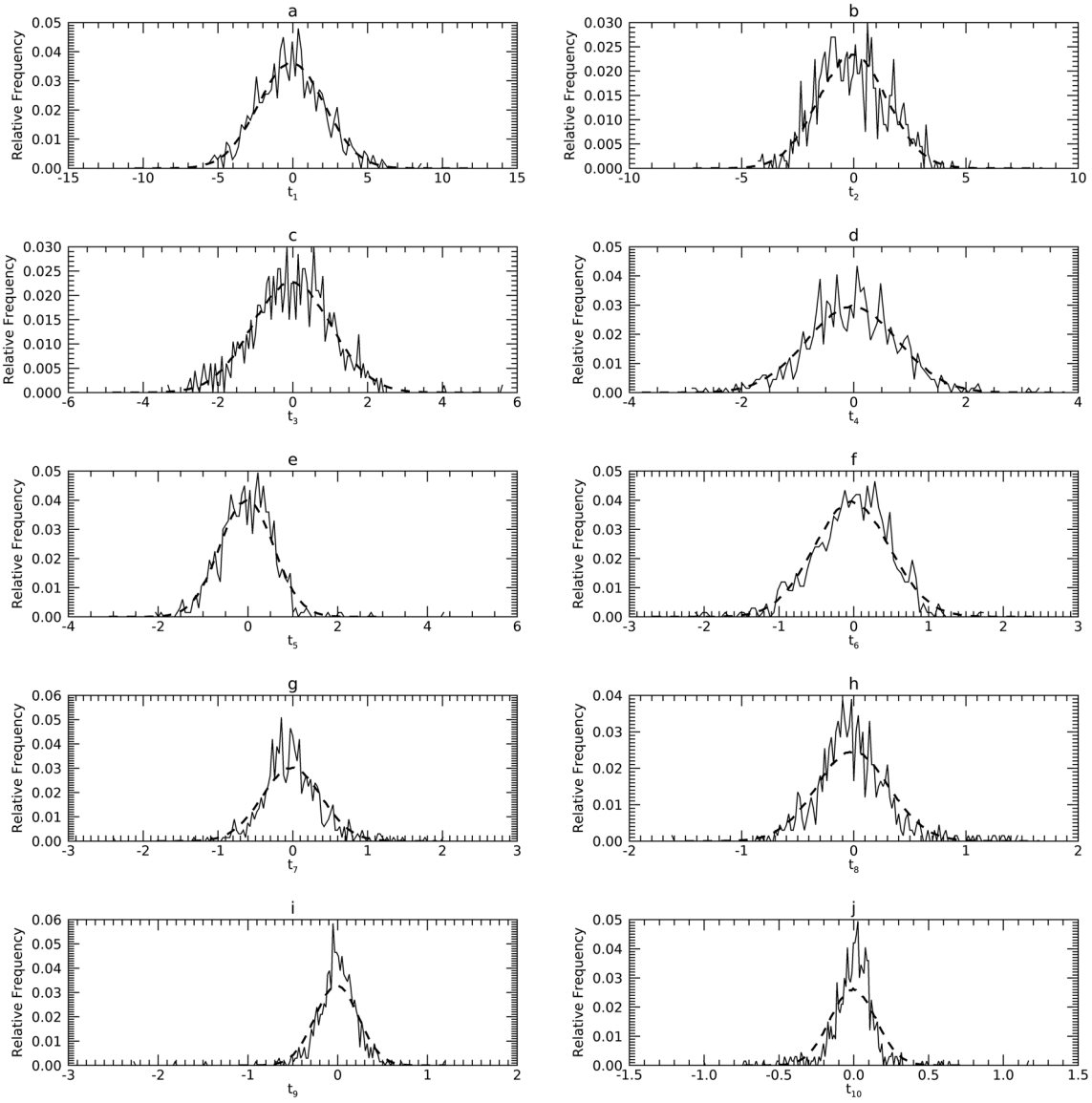

Figure 8. Marginal Probability Density Functions (pdfs) for DS2 in the T representation: each pdf for DS2 (solid) is compared with its corresponding pdf from synthetic data (dashes).

in T when called for with the same method used to augment X for the map/inverse constructions. In this sample scenario, the Y description will deviate from Eq. (1). Although, this premise will have to be investigated because the lack of correlation in T only guarantees t_j independence when $\mathbf{r}(\mathbf{t})$ is multivariate normal. We speculate when the sample has low correlation between most of the bivariate set in X, this approximation may hold.

Choosing the most appropriate space to perform modeling or to analyze the samples deserves consideration. We have used the covariance form suitable for normally distributed variables. In Y, this form is likely appropriate. We used the same form in X as well; this form may not be optimal here because covariance relationships are not preserved over non-linear transformations. It is our contention that Y is best suited for modeling because the

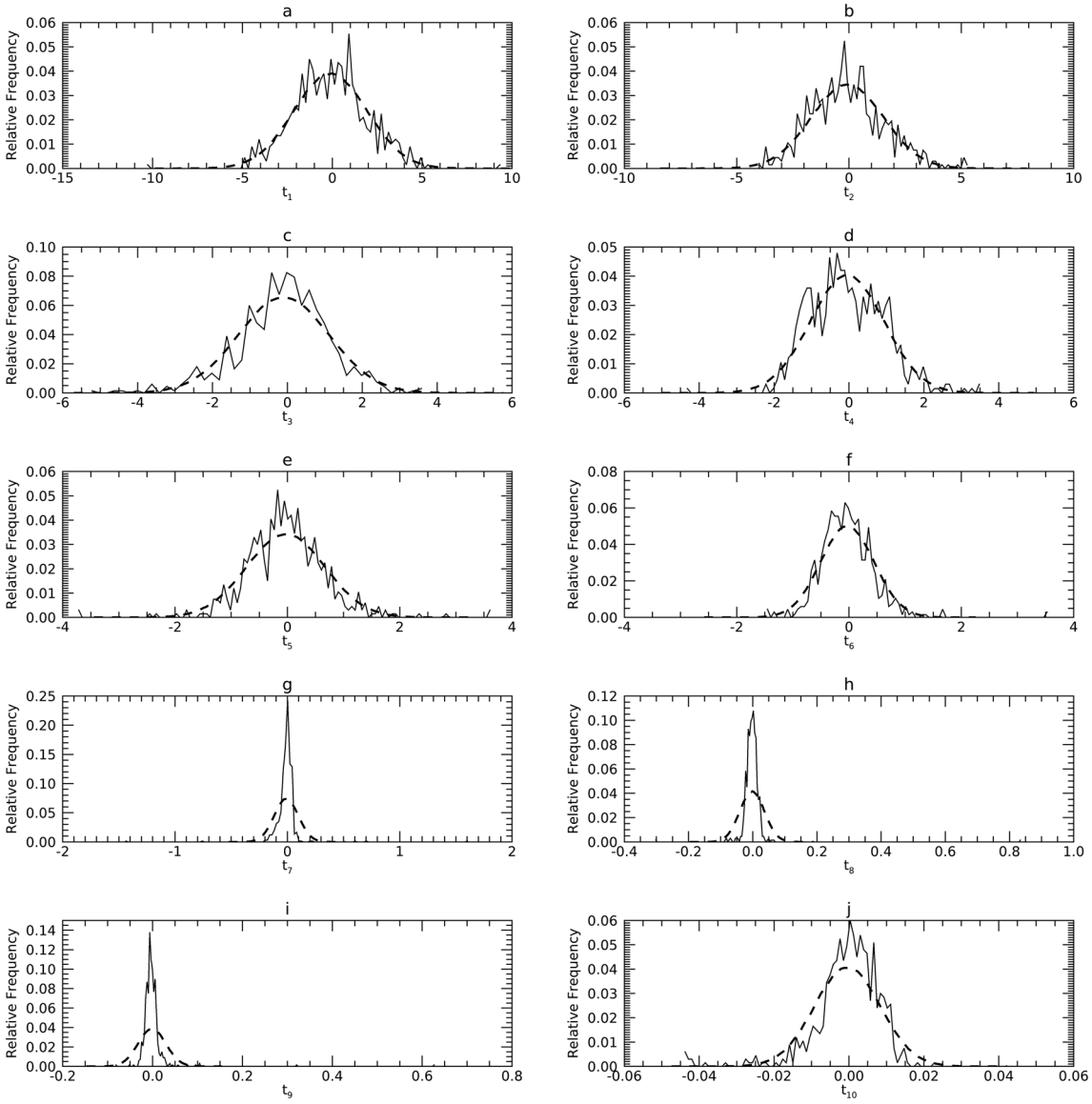

Figure 9. Marginal Probability Density Functions (pdfs) for DS3 in the T representation: each pdf for DS3 (solid) is compared with its corresponding pdf from synthetic data (dashes).

marginals are normal. It is common practice in univariate/multivariate modeling to adjust variables (univariately) to a standardized normal form or apply transforms to remove skewness. The X to Y map converted each x_j to unit variance. When the natural variation for x_j is important, the mapping can be modified easily to preserve the variance. If the variable interpretation is not important, modeling can also be performed in T.

The method in this report addresses the small sample problem given the sample has the latent normal characteristic or is normal. The approach will require further evaluation on different datasets to understand its general applicability and when the univariate mapping from X to Y approximately produces a multivariate normal. Multiple methods were explored to evaluate multivariate normality. These tests indicated that the samples

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
DS1										
Sample (ref)	7.6178	1.0682	0.9634	0.2208	0.0544	0.0244	0.0166	0.0142	0.0116	0.0085
Sample/Syn	7.6142	1.0700	0.9641	0.2213	0.0546	0.0244	0.0166	0.0145	0.0117	0.0087
Syn/Sample	7.6189	1.0987	0.9377	0.2183	0.0520	0.0246	0.0162	0.0138	0.0109	0.0090
DS2										
Sample (ref)	4.8873	2.3806	1.1083	0.6618	0.3575	0.2579	0.1576	0.1064	0.0590	0.0235
Sample/Syn	4.8857	2.3797	1.1072	0.6590	0.3565	0.2567	0.1647	0.1068	0.0596	0.0239
Syn/Sample	4.9237	2.2832	1.1094	0.6908	0.3691	0.2564	0.1714	0.1079	0.0634	0.0248
DS3										
Sample (ref)	4.1932	2.6215	1.4759	0.9790	0.4877	0.2286	0.0116	0.0015	0.0009	0.0001
Sample/Syn	4.1901	2.6125	1.4802	0.9837	0.4888	0.2294	0.0127	0.0015	0.0010	0.0001
Syn/Sample	4.2638	2.6817	1.3871	0.9368	0.4851	0.2312	0.0118	0.0014	0.0009	0.0001

Table 5. Eigenvalues and comparisons: within each dataset, the upper row (sample) gives the eigenvalues of $\mathbf{C}_y$ (sample) used as the reference (ref) for comparisons. The Sample/Syn (condition 1) rows give the eigenvalues (variances) determined by applying the PCA transform derived from the sample to a synthetic (Syn) sample. Syn/Sample (condition 2) rows give the eigenvalues (variances) determined by applying the PCA transform derived from $\mathbf{C}_y$ from a synthetic sample applied to the sample. An F-test was used to compare Sample/Syn t_j with the reference (ref) t_j and to compare Syn/Sample t_j with the reference t_j. In all tests, the null hypothesis was not rejected (i.e. $p > 0.05$).

	DS 1			DS 2			DS 3		
Variable index	t_i	y_i	x_i	t_i	y_i	x_i	t_i	y_i	x_i
1	99.2 (0)	100 (0)	99.0 (0)	99.8 (51.6)	99.4 (98.6)	99.6 (99.2)	99.5 (99.6)	99.5 (99.4)	99.7 (99.6)
2	98.3 (92.2)	100 (0)	99.6 (0)	89.9 (94.6)	99.6 (0)	99.6 (0)	95.7 (79.6)	99.7 (98.2)	99.6 (97.3)
3	94.5 (90.3)	99.8 (0)	99.9 (0)	99.3 (99.1)	99.4 (64.9)	99.5 (62.8)	83.0 (1.6)	99.2 (97.1)	99.1 (96.5)
4	99.8 (4.0)	99.6 (0)	99.7 (0)	98.6 (78.1)	99.1 (26.9)	99.3 (26.3)	94.8 (10.9)	99.6 (99.3)	99.7 (99.5)
5	85.9 (0)	99.9 (0)	99.6 (0)	98.8 (61.4)	99.4 (0)	99.6 (0)	84.6 (0)	99.6 (98.5)	99.6 (99.2)
6	79.7 (0)	99.6 (0)	99.7 (0)	94.9 (76.1)	99.9 (0)	99.6 (0)	71.3 (64.2)	99.7 (0.7)	99.7 (0.7)
7	97.2 (0)	99.8 (0)	99.8 (0)	79.7 (60.8)	99.8 (32.0)	99.7 (29.1)	0 (0)	98.9 (0)	99.0 (0)
8	95.9 (0)	99.7 (0)	99.7 (0)	72.1 (0.9)	99.6 (47.0)	100 (41.7)	0 (0)	99.0 (27.4)	99.4 (29.8)
9	96.0 (0)	98.3 (87.4)	97.6 (73.5)	68.9 (70.6)	99.8 (88.0)	98.3 (84.4)	0 (0)	99.7 (99.7)	99.6 (99.2)
10	99.2 (0)	99.5 (3.1)	99.4 (0.5)	2.8 (8.7)	99.2 (4.2)	99.7 (3.6)	1.2 (0)	98.1 (0)	98.0 (0)

Table 6. Univariate Kolmogorov Smirnov (KS) tests: in each representation, probability density functions (pdfs) derived from synthetic samples were compared with the respective pdfs derived from the sample for each variable. The KS test was applied 1000 times for each comparison. The number of times the null hypotheses was not rejected for a given test was tabulated as a percentage (all entries are percentages). Due to the experimental arrangements for y_j and t_j, each test is equivalent to testing for normality as well. Parenthetical entries show the results when not using kernel density estimation to supplement the X–Y map constructions.

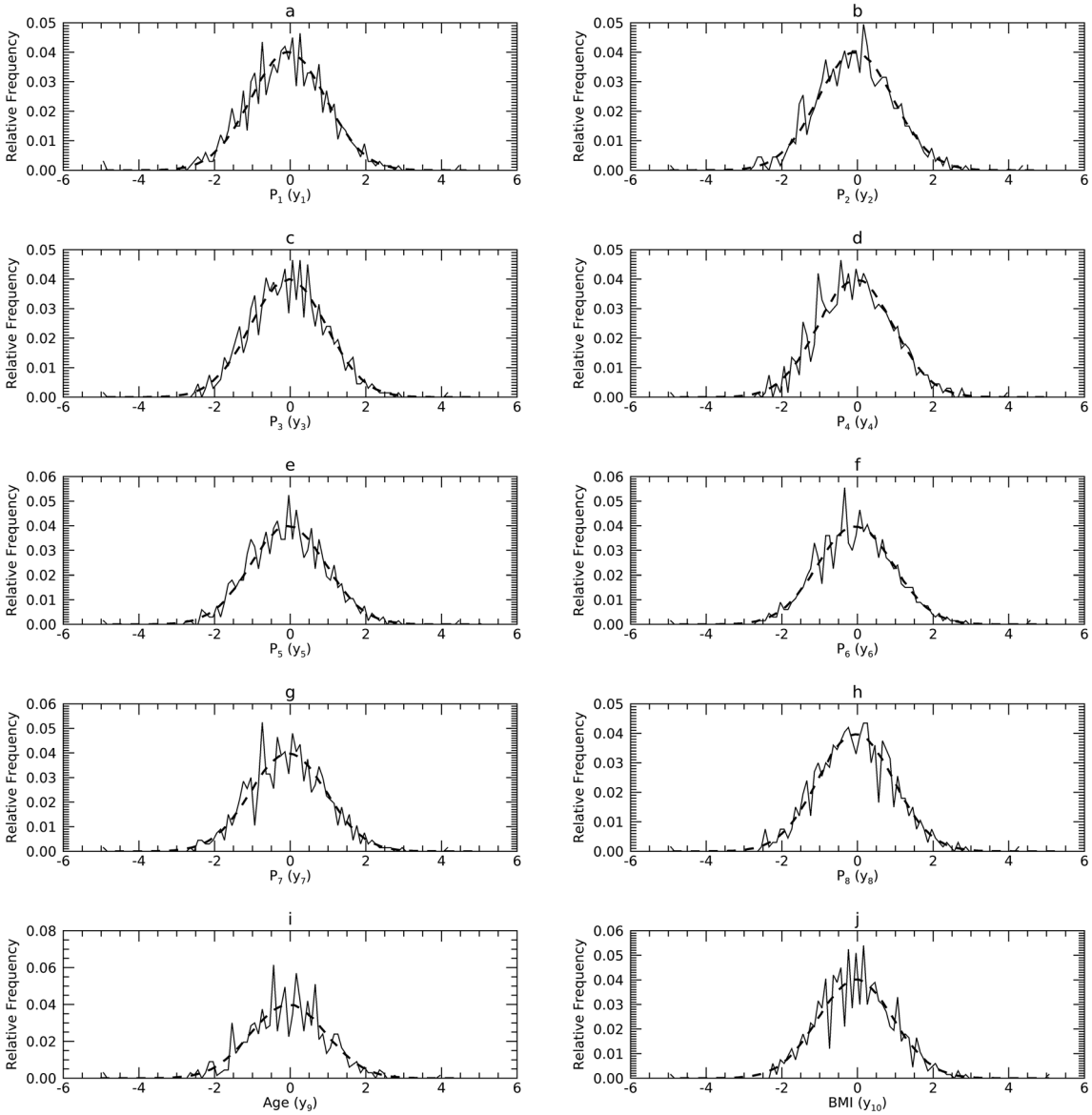

Figure 10. Marginal Probability Density Functions (pdfs) for DS1 in the Y representation: each pdf for DS1 (solid) is compared with its corresponding pdf from synthetic data (dashes). Variables are cited with the names used in their respective resource and with the names used in this report parenthetically.

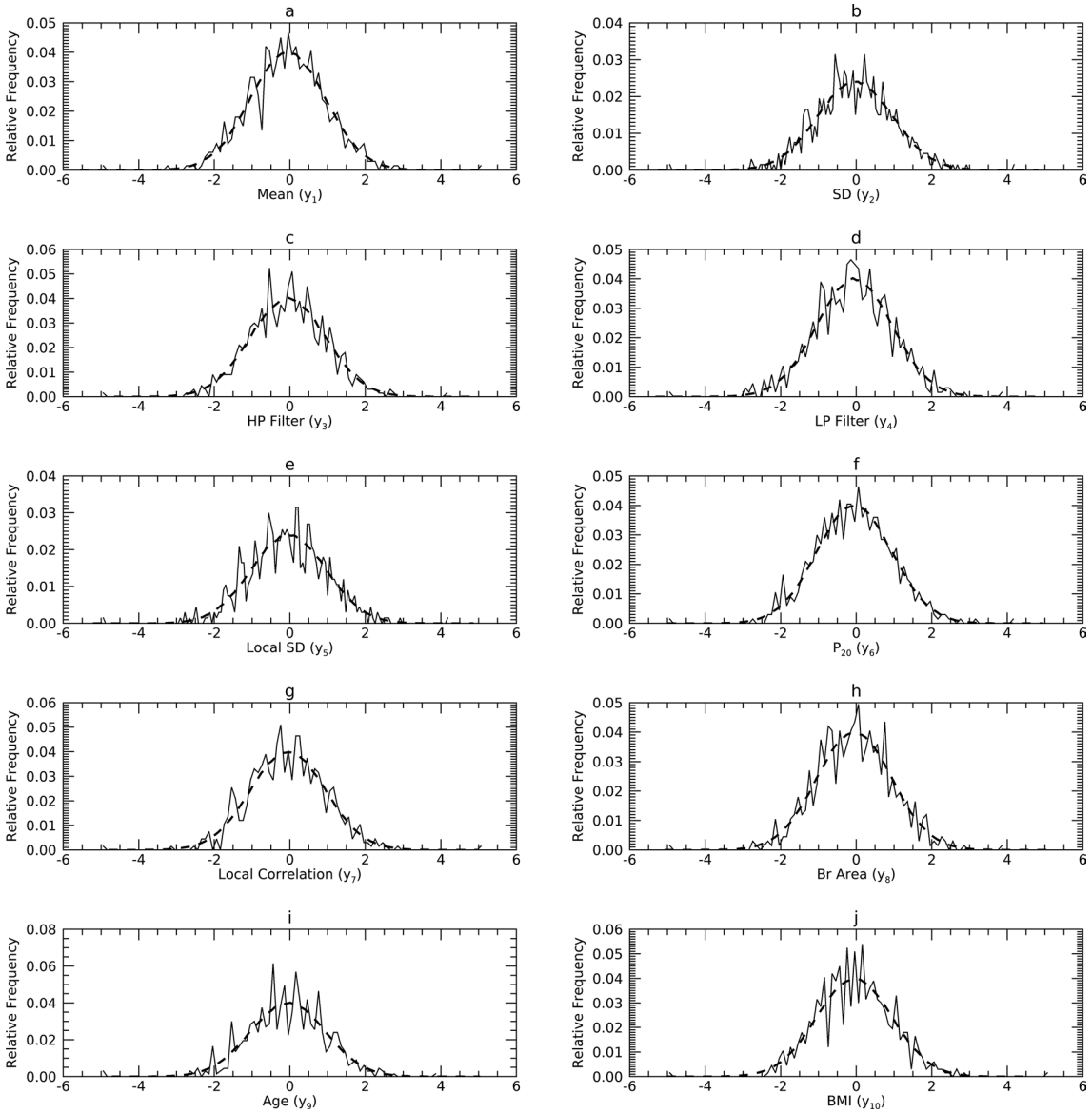

Figure 11. Marginal Probability Density Functions (pdfs) for DS2 in the Y representation: each pdf for DS2 (solid) is compared with its corresponding pdf from synthetic data (dashes). Variables are cited with the names used in their respective resource and with the names used here parenthetically.

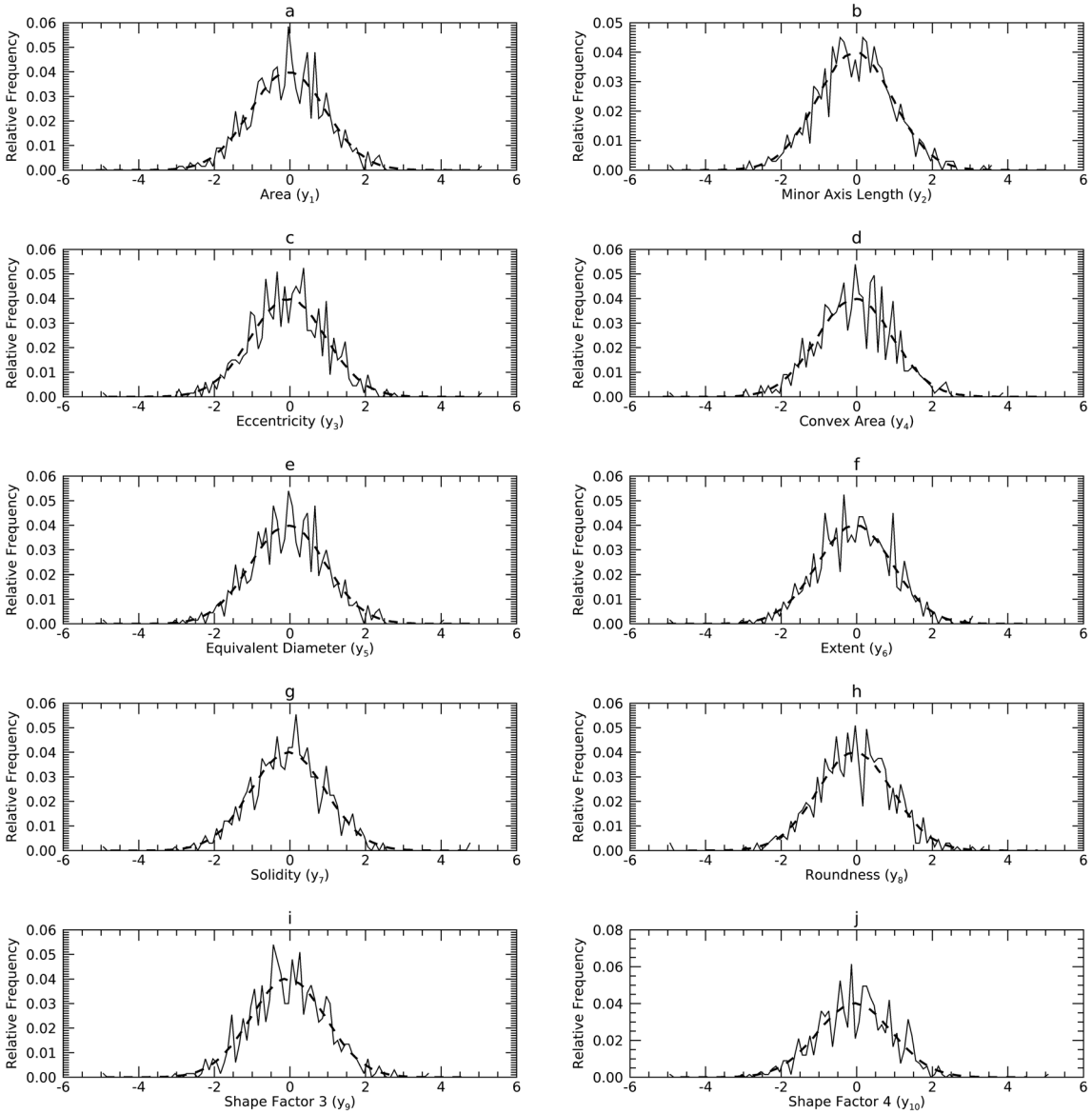

Figure 12. Marginal Probability Density Functions (pdfs) for DS3 in the Y representation: each pdf for DS3 (solid) is compared with its corresponding pdf from synthetic data (dashes). Variables are cited with the names used in their respective resource and with the names used in this report parenthetically.

	MMD test	
x		
DS 1	MMD_c	0.2673
	MMD_u^2	0.0008
DS 2	MMD_c	0.2667
	MMD_u^2	−0.0002
DS 3	MMD_c	0.2681
	MMD_u^2	−0.0001
y		
DS 1	MMD_c	0.2673
	MMD_u^2	0.0005
DS 2	MMD_c	0.2667
	MMD_u^2	−0.0008
DS 3	MMD_c	0.2681
	MMD_u^2	−0.0007
t		
DS 1	MMD_c	0.2673
	MMD_u^2	−0.0005
DS 2	MMD_c	0.2667
	MMD_u^2	−0.0006
DS 3	MMD_c	0.2681
	MMD_u^2	0.0006

Table 7. Multivariate MMD tests: the MMD test was used to compare the sample with 1000 synthetic samples. MMD test quantities are averages over 1000 trials.

approximated normality in both the Y and T but also showed some deviation from normality. The interpretation of these findings in the context of data modeling may aid in understanding the limits of both the multivariate SPs and normality approximations in this report's data and beyond. For example, determining the limiting percentage of the random projection tests may be informative in the modeling context. In summary, we offer a definition for an insufficient sample size in the context of synthetic data. When considering a given sample with d attributes and specified covariance structure, a sample size that does not allow reconstructing its population can be considered as insufficient. In future work, we will apply the methods in this report to understand the minimum sample size, relative to d and a given covariance structure, that permits recovering the population.

		Random projection (%)	Mardia skewness	Mardia kurtosis
x				
DS 1	Sample	40.2	< 0.0001	< 0.0001
	Syn	41.2	< 0.0001	< 0.0001
DS 2	Sample	74.3	< 0.0001	< 0.0001
	Syn	74.1	< 0.0001	< 0.0001
DS 3	Sample	98.9	< 0.0001	< 0.0001
	Syn	99.8	< 0.0001	< 0.0001
y				
DS 1	Sample	98.9	< 0.0001	< 0.0001
	Syn	100.0	0.9445	0.4287
DS 2	Sample	97.4	< 0.0001	< 0.0001
	Syn	99.9	0.4520	0.4374
DS 3	Sample	94.8	< 0.0001	< 0.0001
	Syn	100.0	0.2865	0.0859
t				
DS 1	Sample	99.0	< 0.0001	< 0.0001
	Syn	100.0	0.1620	0.2143
DS 2	Sample	94.1	< 0.0001	< 0.0001
	Syn	100.0	0.8516	0.1933
DS 3	Sample	94.5	< 0.0001	< 0.0001
	Syn	100.0	0.6390	0.4010

Table 8. Multivariate normality tests: for the random projection test, 1000 random projection vectors (1000 tests) were applied to a given sample and each test result was compared against normality; the percentage of times the test was not rejected was tabulated. For synthetic (syn) data, the same 1000 projection vectors were used to test 100 synthetic samples from a given dataset. The percentage of times the test was not rejected was tabulated for each synthetic sample and percentages were averaged over the 100 synthetic samples. Mardia tests were applied to the sample and to a synthetic sample selected at random.

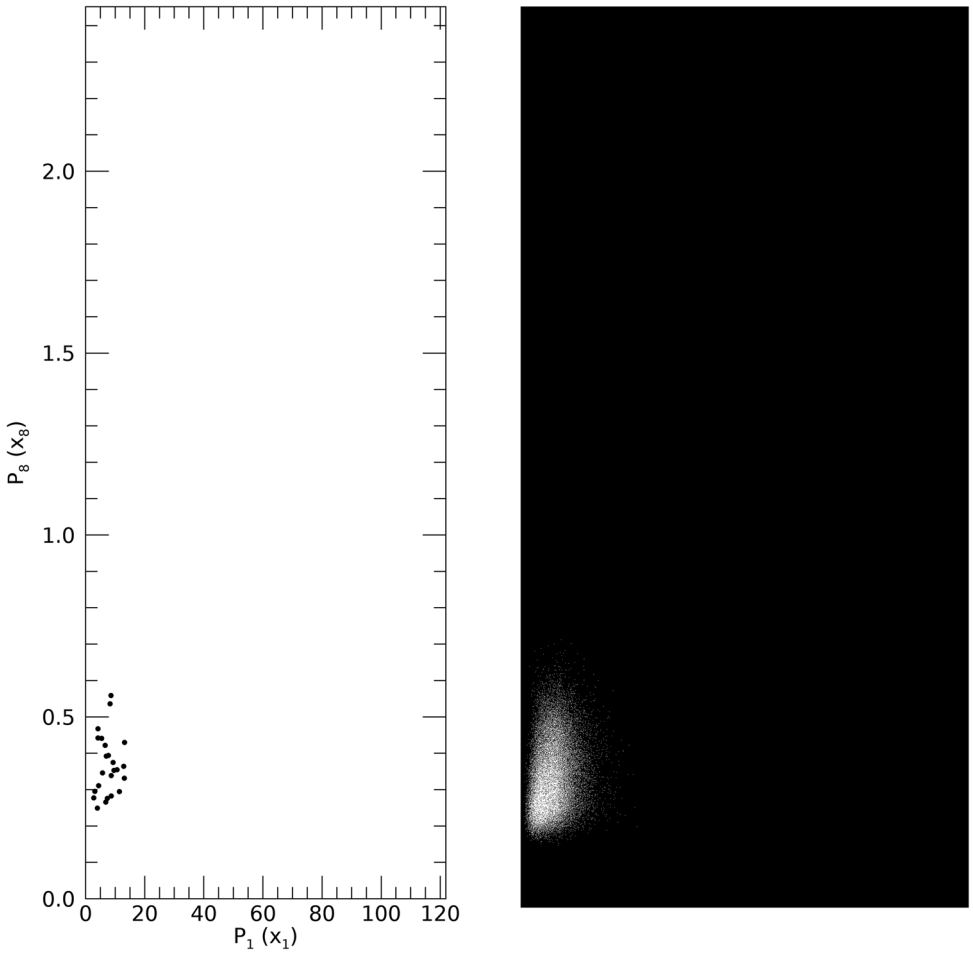

Figure 13. Sample and Synthetic Population Scatter Plot Illustration: a synthetic realization was selected randomly from DS1 giving this measurement vector: $\mathbf{x}^T = [4.20, 2.08, 1.61, 1.15, 0.85, 0.67, 0.54, 0.44, 52.0, 23.6]$. x_1 and x_8 were selected for the scatter plot variables. For the other 7 components, all synthetic realizations within ± ½ (the respective standard deviation) were selected and viewed as a scatter plot in the x_1x_8 plane. Using the same vector and limits, individuals were selected from the sample in the same manner. The scatter plot for the sample (n = 24) is shown in the left-pane and from the synthetic population (n = 36,398) in the right-pane. Variables are cited with the names used in their respective resource and with the names used in this report parenthetically.

Data availability

The link to publicly available data is provided in text. Mammography summary data can be obtained upon request to the corresponding author: John Heine (john.heine@moffitt.org). Kernel parameters are also available upon request.

Received: 6 April 2023; Accepted: 16 July 2023
Published online: 28 July 2023

References

1. Gail, M. H. & Pfeiffer, R. M. Breast cancer risk model requirements for counseling, prevention, and screening. *J. Natl. Cancer Inst.* **110**, 994–1002 (2018).
2. Garrido-Castro, A. C. & Winer, E. P. Predicting breast cancer therapeutic response. *Nat. Med.* **24**, 535–537 (2018).
3. Huo, Z. *et al.* Automated computerized classification of malignant and benign masses on digitized mammograms. *Acad. Radiol.* **5**, 155–168 (1998).
4. Lei, C. *et al.* Mammography-based radiomic analysis for predicting benign BI-RADS category 4 calcifications. *Eur. J. Radiol.* **121**, 108711. https://doi.org/10.1016/j.ejrad.2019.108711 (2019).
5. Nguyen, D. V. & Rocke, D. M. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18**, 39–50 (2002).
6. Erves, J. C. *et al.* Needs, priorities, and recommendations for engaging underrepresented populations in clinical research: A community perspective. *J. Community Health* **42**, 472–480. https://doi.org/10.1007/s10900-016-0279-2 (2017).
7. Dickson, J. L. *et al.* Hesitancy around low-dose CT screening for lung cancer. *Ann. Oncol.* **33**, 34–41. https://doi.org/10.1016/j.annonc.2021.09.008 (2022).

8. Wang, G. X. *et al.* Barriers to lung cancer screening engagement from the patient and provider perspective. *Radiology* **290**, 278–287. https://doi.org/10.1148/radiol.2018180212 (2019).
9. Foraker, R., Mann, D. L. & Payne, P. R. O. Are synthetic data derivatives the future of translational medicine?. *JACC Basic Transl. Sci.* **3**, 716–718 (2018).
10. Elston, D. M. Data dredging and false discovery. *J. Am. Acad. Dermatol.* **82**, 1301–1302. https://doi.org/10.1016/j.jaad.2019.07.061 (2020).
11. Siddiqui, K. Heuristics for sample size determination in multivariate statistical techniques. *World Appl. Sci. J.* **27**, 285–287 (2013).
12. Wu, Y., Genton, M. G. & Stefanski, L. A. A multivariate two-sample mean test for small sample size and missing data. *Biometrics* **62**, 877–885 (2006).
13. Riley, R. D. *et al.* Calculating the sample size required for developing a clinical prediction model. *BMJ* **368**, m441. https://doi.org/10.1136/bmj.m441 (2020).
14. Geladi, P. & Kowalski, B. R. Partial least-squares regression: A tutorial. *Anal. Chim.* **185**, 1–17 (1986).
15. Chartrand, G. *et al.* Deep learning: A primer for radiologists. *Radiographics* **37**, 2113–2131 (2017).
16. Buczak, A. L., Babin, S. & Moniz, L. Data-driven approach for creating synthetic electronic medical records. *BMC Med. Inform. Decis.* **10**, 1–28 (2010).
17. Chen, J. Q., Chun, D., Patel, M., Chiang, E. & James, J. The validity of synthetic clinical data: A validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med. Inform. Decis. Mak.* https://doi.org/10.1186/s12911-019-0793-0 (2019).
18. Dahmen, J. & Cook, D. A synthetic data generation system for healthcare applications. *Sensors (Basel)* **19**, 1181. https://doi.org/10.3390/s19051181 (2019).
19. Goncalves, A. R., Sales, A. P., Ray, P. & Soper, B. NCI pilot 3-synthetic data generation report report no. Lawrence Livermore National Lab. (LLNL): LLNL-TR-747902 (2018).
20. Bogle, B. M. & Mehrotra, S. A moment matching approach for generating synthetic data. *Big Data* **4**, 160–178 (2016).
21. Quintana, D. S. A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *Elife* **9**, e53275 (2020).
22. Fowler, E. E., Berglund, A., Sellers, T. A., Eschrich, S. & Heine, J. Empirically-derived synthetic populations to mitigate small sample sizes. *J. Biomed. Inform.* **105**, 103408 (2020).
23. Scott, D. W. Feasibility of multivariate density estimates. *Biometrika* **78**, 197–205 (1991).
24. Hwang, J.-N., Lay, S.-R. & Lippman, A. Nonparametric multivariate density estimation: A comparative study. *IEEE Trans. Signal Process.* **42**, 2795–2810 (1994).
25. Wang, Z. & Scott, D. W. Nonparametric density estimation for high-dimensional data—Algorithms and applications. *Wiley Interdiscip. Rev. Comput. Stat.* **11**, e1461 (2019).
26. Heine, J., Fowler, E. E. E., Berglund, A., Schell, M. J. & Eschrich, S. A. Techniques to produce and evaluate realistic multivariate synthetic data. *bioRxiv.* https://doi.org/10.1101/2021.10.26.465952 (2021).
27. Price, K. V., Storn, R. M. & Lampinen, J. A. *Differential Evolution: A Practical Approach to Global Optimization* (Springer, 2005).
28. Koklu, M. & Ozkan, I. A. Multiclass classification of dry beans using computer vision and machine learning techniques. *Comput. Electron. Agric.* **174**, 105507 (2020).
29. Fowler, E. E. E. *et al.* Generalized breast density metrics. *Phys. Med. Biol.* **64**, 015006. https://doi.org/10.1088/1361-6560/aaf307 (2019).
30. Heine, J. J. & Velthuizen, R. P. Spectral analysis of full field digital mammography data. *Med. Phys.* **29**, 647–661 (2002).
31. Fowler, E. E. E. *et al.* Spatial correlation and breast cancer risk. *Biomed. Phys. Eng. Express* **5**, 045007. https://doi.org/10.1088/2057-1976/ab1dad (2019).
32. Press, W. H., Numerical Recipes Software (Firm). *Numerical Recipes in C* 2nd edn. (Cambridge University Press, 1992).
33. Oh, H. *et al.* Early-Life and adult anthropometrics in relation to mammographic image intensity variation in the nurses' health studies. *Cancer Epidemiol. Biomark. Prev.* **29**, 343–351. https://doi.org/10.1158/1055-9965.EPI-19-0832 (2020).
34. Velthuzen, R. P. & Clarke, L. P. In *SPIE proceedings series.* 179–187 (Society of Photo-Optical Instrumentation Engineers).
35. Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.* **13**, 723–773. https://doi.org/10.5555/2188385.2188410 (2012).
36. Garreau, D., Jitkrittum, W. & Kanagawa, M. Large sample analysis of the median heuristic. arXiv preprint https://arxiv.org/abs/1707.07269 (2017).
37. Zhou, M. & Shao, Y. A powerful test for multivariate normality. *J. Appl. Stat.* **41**, 351–363. https://doi.org/10.1080/02664763.2013.839637 (2014).
38. Shao, Y. & Zhou, M. A characterization of multivariate normality through univariate projections. *J. Multivar. Anal.* **101**, 2637–2640. https://doi.org/10.1016/j.jmva.2010.04.015 (2010).
39. Haugh, M. An introduction to copulas. In *IEOR E4602: Quantitative Risk Management. Lecture Notes* (Columbia University, 2016).
40. Durante, F., Fernández-Sánchez, J. & Sempi, C. *Aggregation Functions in Theory and in Practise* 85–90 (Springer, 2013).
41. Schirmacher, D. & Schirmacher, E. *Multivariate Dependence Modeling Using Pair-Copulas* (The Society of Actuaries, 2008).
42. Chandrasekara, N. & Tilakaratne, C. D. Determining and comparing multivariate distributions: An application to AORD and GSPC with their related financial markets. *GSTF J. Math. Stat. Oper. Res. JMSOR* **4**, 1–8 (2016).
43. Jones, M. C., Marron, J. S. & Sheather, S. J. A brief survey of bandwidth selection for density estimation. *J. Am. Stat. Assoc.* **91**, 401–407 (1996).
44. Gramacki, A. *Nonparametric Kernel Density Estimation and Its Computational Aspects* (Springer, Berlin, 2018).
45. Schrab, A. *et al.* MMD aggregated two-sample test. arXiv preprint https://arxiv.org/abs/2110.15073 (2021).
46. Korkmaz, S., Göksülük, D. & Zararsiz, G. MVN: An R package for assessing multivariate normality. *R J.* **6**, 151 (2014).
47. Farrell, P. J., Salibian-Barrera, M. & Naczk, K. On tests for multivariate normality and associated simulation studies. *J. Stat. Comput. Simul.* **77**, 1065–1080 (2007).

Author contributions

J.H. is the corresponding author, conceived the plan and methods; E.F. is a coauthor, developed the computer code, assisted in the plan and methods development, and prepared figures; A.B. is a coauthor and provided statistical and principal component analysis expertise; M.S. is a coauthor and provided statistical expertise; S.E. is a coauthor and assisted in the plan and methods developments. All authors reviewed the manuscript.

Funding

The work was in part supported by Moffitt Cancer Center grant #17032001 (Miles for Moffitt) and National Institutes of Health Grants R01CA166269 and U01CA200464.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to J.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.