MarketVision RESEARCH®

SYNTHETIC DATA PERSPECTIVE:

THE PROMISE & LIMITATIONS OF SYNTHETIC DATA FOR QUANTITATIVE RESEARCH

Synthetic Data Has Many Applications

Synthetic data, generated using AI/machine learning (ML) models, is intended to mimic the statistical structure of real-world data. The term "synthetic data" encompasses a wide variety of emerging techniques used in both qualitative and quantitative methodologies. These range from segment personas, digital twins, and sample augmentation to applications that support exploratory analysis, stress-testing, or sensitivity analysis, and simulation of concept or message performance under different assumptions.

Even within the specific use case of generating responses for quantitative surveys, methodologies can vary significantly. Some implementations use Large Language Models (LLMs) to "answer" questions like humans, while others use ML models that statistically learn and replicate dataset structures.

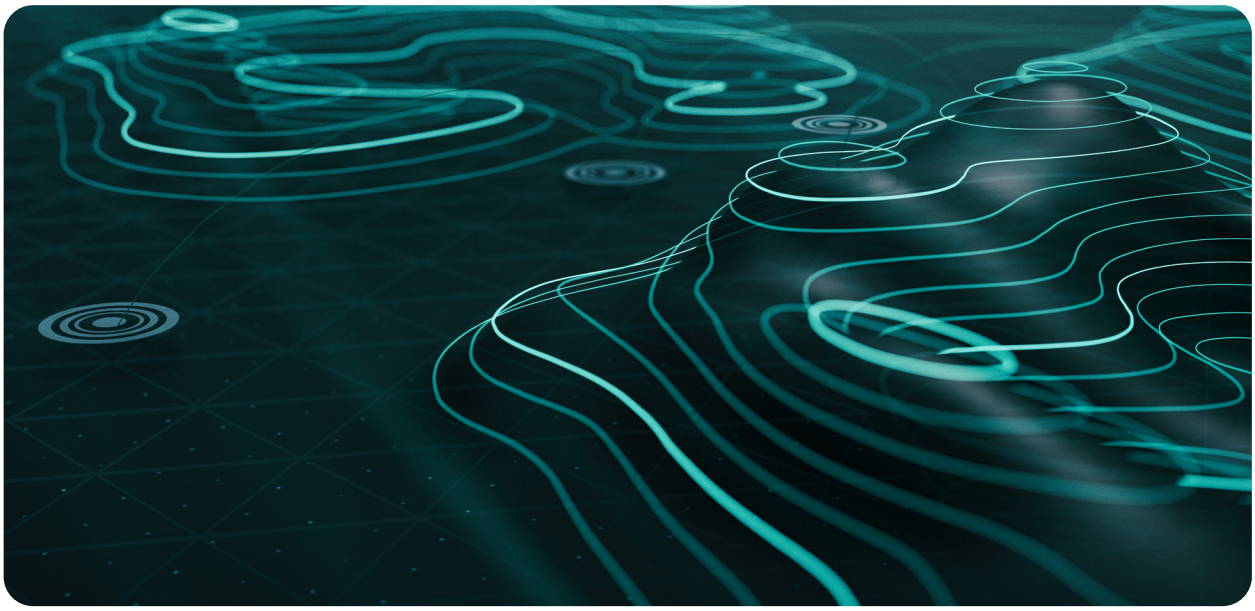
LLM-based approaches generally depend on prompting or conditioning strategies to approximate how a member of a target group might respond, whereas purpose-built ML models are trained directly on empirical survey data from the population of interest. As a result, the two approaches differ in the degree to which population-specific patterns are explicitly learned versus inferred, which can materially affect reliability in quantitative research settings.

Synthetic data shows some promise for researchers, but caution is needed

Synthetic data applications are rapidly evolving, which is the nature of uncharted territory. It's like mapping an unexplored "synthetic continent". Everyone agrees that the coastlines are clearly visible (i.e., it's AI-generated and not real-world data), but the inner mainland (i.e., viable applications) remains uncertain. As that investigation moves forward, there are a growing number of third-party providers offering synthetic data who have invested significant time and resources developing solutions with varied approaches and levels of evidence for their marketing claims.

There is reason to be optimistic about targeted use of synthetic data in scenarios where the underlying dataset is large enough to support meaningful ML model training. One of the most promising applications of synthetic data is the ability to bolster niche subgroups within a larger sample by simulating additional members of a cohort that would otherwise be too small to analyze statistically. When applied correctly, meaning in ways that support exploratory analysis and scenario development rather than replace primary data, synthetic data may support more confident evaluations of segments, exploratory modeling, and help generate 'what-if' scenarios or hypotheses for testing with traditional methodologies.

There is also reason to be cautious. While synthetic data offers potential cost and time savings, it must be grounded in enough high-quality, non-synthetic data to be reliable. The question of what constitutes "enough" is still very much under debate. Different modeling approaches have different data thresholds, and vendors vary in their claims about minimum sample requirements. What represents appropriate use or sufficient validation in one research context may be inadequate in another, as expectations can differ substantially across categories, audiences, and decision stakes.



Synthetic data must be validated to ensure complex structural relationships are preserved

Synthetic data cannot be assumed to replicate the full complexity of consumer experience without appropriate and context-specific validation that addresses the structure of the data in comparison to the real-world data. This type of validation is not a one-time exercise, but an ongoing practice that is critical in small-sample environments, when the desired data analysis relies on the interrelationships of variables, and especially when the risk of relying on inaccurate data is significant.

Ideally, synthetic data of the highest quality could simulate real-world multivariate modeling results, even if it is not able to reproduce output identically. The critical factor to assess is the degree to which it preserves the underlying data structure and patterns closely enough that deviations either bring to light linkages that can be inferred from the existing relationships among variables, or are within an acceptable error range. To date, our results show that open-source synthetic augmentation of "small" samples (e.g., n=300-600, as are commonly used in market research due to budget constraints and recruiting feasibility) is more likely to introduce structural inconsistencies than meaningful extensions of the data.

Evaluating synthetic data requires looking beyond high-level metrics to assess the model's ability to generalize effectively instead of simply replicating training data. To be useful, synthetic data must preserve realistic variance, covariance, and correlations between responses, along with simple means and univariate distributions. Poorly trained models can generate records that appear acceptable on the surface but upon deeper inspection include discrepancies that can have a compound effect across the entire dataset, creating cumulative errors that distort insights to a meaningful and potentially statistically significant degree. In most applied settings, this means it is critical to have sufficient real data available to support situational validation of synthetic outputs, as performance can vary substantially across use cases and populations.

Any failure to preserve the true complexity of the data means advanced analyses will not accurately model the consumer experience, leading stakeholders to make conclusions and decisions that are not based on reality, yet appear to be data-driven. A key hurdle for this technology to overcome will be the tendency for synthetic data to "smooth out" true population variance, underrepresent strong reactions, and overrepresent the average, potentially making statistical tests look more significant and leading to "confidently incorrect" conclusions.

Navigating the frontier – Balancing potential and statistical reality

While the "coastline" of the synthetic data landscape is increasingly visible, the reliability of its application depends heavily on the volume and availability of high-quality empirical survey data for training and validation. The reliability of any synthetic application is closely tied to the quality of the empirical evidence used to build it. This creates a functional paradox for researchers: synthetic data is often most desired when real-world data is scarce, yet machine learning models require substantial real-world data to learn accurately. Without adequate grounding in the observed data, models tend to drift toward generic patterns, losing the nuance that differentiates a specific target audience from the general population. Therefore, the "inland" territory of synthetic data is best navigated not as a way to generate data from nothing, but as a method to amplify and stress-test relationships that have already been firmly established by robust primary research.

We believe that synthetic data, while a promising frontier for market research, is not currently a universal substitute for primary data collection. Researchers must exercise rigorous, context-specific validation beyond simple distribution checks. Synthetic data must be treated not as an independent source of truth, but as sophisticated simulations that require human oversight to ensure they support rather than distort, data-driven decision-making.