Evaluating Synthetic Survey Estimates

Reducing Noise While Preserving Opinion Structure

Executive Summary

Survey research is a powerful tool for understanding opinions and behaviors, but live survey results are not synonymous with ground truth. In modern survey methodology, observed estimates are understood as measurements subject to multiple sources of error rather than direct representations of underlying population parameters (8; 9).

This paper presents a validation framework for evaluating synthetic survey estimates as *model-based estimators* of opinion structure rather than literal reproductions of a single live sample. Using a physician sarcopenia study as a case example, we demonstrate that synthetic estimates can, in some contexts, reduce sampling noise and improve internal consistency while preserving clinically plausible patterns. The claims advanced here are deliberately conservative: synthetic data does not reveal objective truth, but it may approximate expected distributions under known constraints with measurable fidelity.

Contents

1 Surveys as Observations, Not Truth

1.1 Total Survey Error

Modern survey methodology conceptualizes error through the Total Survey Error (TSE) framework, which recognizes that deviations between survey estimates and population parameters arise from multiple sources, including sampling error, coverage error, nonresponse error, and measurement error (8; 2). Even when surveys are carefully designed, finite sample sizes introduce variability that can materially affect reported results, particularly in healthcare studies with typical sample sizes of a few hundred respondents.

Under this framework, a live survey represents one realization of a stochastic measurement process rather than a definitive benchmark. Consequently, divergence between two estimates does not, by itself, imply error in either; it may instead reflect expected variability inherent to survey measurement (9).

1.2 Respondent Behavior and Measurement Error

In addition to sampling variability, survey estimates are influenced by respondent behavior. A substantial body of research documents that respondents frequently engage in satisficing—adopting cognitively economical response strategies—especially in long instruments, grid formats, and multi-select questions (11; 12; 19). These behaviors can systematically distort marginal distributions without reflecting true underlying preferences.

Multi-select questions are particularly susceptible to such effects, as respondents may under-select options, favor familiar items, or terminate consideration early (17). Grid questions similarly encourage satisficing through central tendency bias and straight-lining (20). These measurement effects introduce structured noise that is difficult to disentangle from signal when relying on a single live sample.

2 Synthetic Estimates as Model-Based Estimators

2.1 Borrowing Strength and Noise Reduction

Model-based approaches trained on large, related datasets can borrow strength across correlated variables, a principle closely related to shrinkage and partial pooling in statistical estimation. Classical and Bayesian results demonstrate that such estimators can reduce expected mean squared error by trading small increases in bias for substantial reductions in variance (10; 5; 6).

From this perspective, differences between a synthetic estimate and a single live survey realization may reflect noise reduction rather than model misspecification, particularly when the live sample

is small or noisy.

2.2 Relation to Accepted Survey Methods

The use of models to improve survey estimates has a well-established precedent in survey research. Techniques such as multilevel regression and poststratification (MRP) combine observed survey responses with population structure to produce more stable estimates, especially for small subgroups (7; 15; 16).

Synthetic survey estimation can be viewed as an extension of this paradigm, in which learned relationships from historical data are combined with explicit study constraints to estimate expected response distributions.

3 Evaluation Framework

3.1 Question-Type–Aware Metrics

Because different survey question formats encode information differently, no single evaluation metric is appropriate for all questions. Validation must therefore be sensitive to question type rather than relying on a uniform summary statistic (18).

3.2 Distributional Similarity for Single-Select Questions

For single-select and ordinal questions, responses form a probability distribution over mutually exclusive categories. Distributional similarity metrics, such as Kullback–Leibler (KL) divergence, provide a principled measure of how closely a simulated distribution approximates an observed one when the live survey is treated as the reference distribution (13).

KL divergence quantifies the information lost when using the synthetic distribution to approximate the live distribution. Lower values indicate closer correspondence, with 0 representing perfect match. In practice, KL values below 0.15 indicate excellent agreement, while values between 0.15 and 0.30 represent good agreement with minor distributional differences.

3.3 Rank Similarity for Multi-Select Questions

Multi-select questions do not constitute probability distributions, as each response option is selected independently. In these cases, rank-based measures better capture realism by emphasizing the relative ordering of response options.

We use Rank-Biased Overlap (RBO) with parameter $p = 0.9$, which provides top-weighted similarity aligned with expert qualitative review (22). This weighting reflects the practical reality that

agreement on the most-selected options is more important than exact ordering of rarely-chosen items. RBO ranges from 0 to 1, where higher values indicate stronger ranking agreement, with emphasis on top-ranked items. RBO values above 0.70 indicate good to excellent ranking agreement.

4 Case Study: Physician Sarcopenia Survey

4.1 Study Design

A live physician survey on sarcopenia ($n = 253$) was replicated using multiple synthetic simulation strategies with varying constraints, including specialty quotas, targeting rules, and sample size. The primary comparison presented here focuses on the validation run (Run 6, $n = 1000$), which used a mixed approach combining explicit specialty quotas (27% Family Medicine, 27% Internal Medicine) with targeting rules (16% Geriatric Medicine, 30% Physical Medicine & Rehabilitation) to exactly mirror the live study's specialty composition.

All comparisons treat the live survey as the reference observation and evaluate the synthetic results as alternative estimates of the same underlying opinion structure.

4.2 Distributional Similarity: Single-Select Questions

Across single-select and ordinal questions, similarity between live and synthetic results was evaluated using KL divergence. KL values were generally low to moderate, indicating close correspondence between the simulated and observed distributions.

Rather than reproducing every marginal proportion exactly, the synthetic estimates tended to smooth extreme values observed in the live sample while preserving overall shape and directional patterns. This behavior is consistent with reduced sampling variance in the larger synthetic sample and with model-based partial pooling effects discussed in Section 2.

4.3 Rank Similarity: Multi-Select Questions

Multi-select questions were evaluated using rank-based similarity. For these items, response options were ordered by selection frequency in both the live and synthetic data, and RBO was used to assess similarity between rankings.

Across multi-select items, rank similarity was generally strong, with the highest-priority options showing the greatest agreement. Differences between live and synthetic rankings were concentrated among mid- and lower-ranked options, which are known to be particularly sensitive to respondent satisficing and sample composition.

4.4 Summary of Validation Outcomes

Table 1 summarizes overall validation performance across question types using heuristic similarity thresholds. These thresholds (KL ≤ 0.15 for excellent agreement on single-select questions; RBO ≥ 0.70 for good ranking agreement on multi-select questions) represent practical benchmarks rather than formal statistical tests.

Table 1: Summary of Validation Outcomes

Question Type	**Total**	**Excellent/Good**	**Moderate**	**Notes**
Single-select (KL ≤ 0.15)	29	16 (55%)	11 (38%)	2 questions show higher divergence but preserve clinical logic
Multi-select (RBO ≥ 0.70)	10	9 (90%)	1 (10%)	Strong top-ranking agreement across questions

Note: RBO emphasizes agreement on most-selected options, reflecting the practical insight that "getting the top 3-5 priorities right" matters more than exact ordering of less-important items.

Across single-select questions, a majority exhibited low to moderate KL divergence, while a smaller subset showed higher divergence, typically associated with polarized responses or questions where synthetic estimates diverged in clinically interpretable ways. For multi-select questions, rank similarity was strong in nearly all cases, with agreement concentrated among top-ranked options.

4.5 Illustrative Validation Examples

To ground the validation framework in concrete evidence, we present two representative examples that demonstrate different aspects of synthetic data performance.

4.5.1 Example 1: Single-Select Comparison (Distributional Agreement)

Table 2 presents live and synthetic response distributions for a representative ordinal grid item from Q16 assessing whether patients with sarcopenia have "low physical activity" as a characteristic.

This grid item demonstrates excellent distributional agreement (KL = 0.081), with both live and synthetic responses concentrating on "Often" and "Almost always" categories. The dominant response pattern—that low physical activity frequently applies to sarcopenia patients—is preserved in the synthetic estimate. Minor differences in the "Sometimes" and "Almost always" categories likely reflect sampling variance rather than substantive disagreement, illustrating how synthetic estimates can capture clinical consensus while smoothing extreme probability mass concentrations.

Table 2: Live vs. Synthetic Comparison for Q16 Grid Item (Low Physical Activity)

Response Option	**Live (%)**	**Synthetic (%)**
Never	0	0
Rarely	1	1
Sometimes	7	18
Often	49	48
Almost always	43	32
N/A – Not enough experience	0	2
KL Divergence		**0.081**

4.5.2 Example 2: Multi-Select Comparison (Preserving Priority Structure)

Table 3 presents a ranked comparison for the multi-select question assessing primary motivations to screen and treat sarcopenia (Q24), where physicians could select up to two options.

Table 3: Live vs. Synthetic Comparison for Multi-Select Question (Q24 - Primary Motivations)

Rank	**Response Option**	**Live (%)**	**Synthetic (%)**
1	Fall/injury prevention	60	41
2	Concern for ability to remain independent/mobile	53	51
3	Concern for decreased strength during ADLs	26	40
4	Concern for worsening comorbidities	15	35
5	Concern for higher risk of mortality	23	33
6	Compliance with Medicare regulations	4	16
7	Patient or caregiver satisfaction	8	22
RBO ($p = 0.9$)			**0.882**

The rankings show strong agreement (RBO = 0.882), with fall prevention and maintaining independence consistently identified as top priorities in both live and synthetic responses. The live sample shows extreme concentration on the top item (60%), while synthetic estimates distribute support more evenly across top motivations. This pattern may reflect reduced satisficing from "select up to two" constraints, where respondents in live surveys disproportionately select the first satisfactory option rather than considering all alternatives (17). The preserved ranking structure demonstrates that synthetic data successfully captures the priority hierarchy that matters most for clinical decision-making.

4.6 When Synthetic Estimates May Provide Additional Value

In several instances, synthetic estimates diverged from live results in ways that may reduce measurement error or better represent guideline-adherent practice. These divergences, while resulting in higher KL values or moderate ranking differences, may provide value beyond simple validation.

4.6.1 Reducing Satisficing in Follow-Up Assessment (Q15)

For the question "Once a patient has been diagnosed, how often do you perform follow-up assessment?", the live survey showed 64% of physicians reporting they assess patients "every visit" after diagnosis, while synthetic estimates showed 34% (KL = 0.208). This divergence may reflect known satisficing patterns in clinical practice questions, where respondents select the most socially desirable response (11). The synthetic distribution shows more realistic variance across all response options (once a year: 17% live vs. 25% synthetic; only if patient complains: 7% live vs. 14% synthetic), consistent with reducing measurement error while preserving the insight that regular follow-up is common practice.

4.6.2 Capturing Guideline-Adherent Screening Practices (Q9)

Multi-select question Q9 asked about screening tools and measures typically used. While ranking similarity remained good (RBO = 0.673), synthetic estimates showed substantially higher adoption of evidence-based assessment tools:

- SPPB test: 6% (live) vs. 33% (synthetic)
- SARC-F questionnaire: 4% (live) vs. 34% (synthetic)
- DEXA scan: 16% (live) vs. 47% (synthetic)

Importantly, the overall ranking structure was preserved—simple functional tests like "get up and go" (65% live, 47% synthetic) and grip strength (36% live, 47% synthetic) remained among the most popular in both distributions. The divergence may reflect synthetic data's tendency toward guideline-following practice patterns (3), potentially useful for modeling optimal clinical pathways rather than current real-world adoption rates. This distinction is valuable when the research objective is understanding what engaged, protocol-adherent physicians do, rather than measuring actual adoption rates across all practice styles.

4.6.3 Professional Proactivity in Initial Screening (Q8)

For the question "Which of the following people is most likely to express initial concern?" about sarcopenia, synthetic physicians were more likely to identify themselves or their care team as initiating screening (33% vs. 11%), while live physicians more often reported family members raising initial concerns (60% vs. 28%). This divergence may reflect synthetic data capturing engaged, proactive clinical practice rather than reactive response to external prompts—a distinction valuable for understanding optimal versus typical care pathways. When designing interventions or care protocols, knowing what proactive physicians do may be more useful than knowing what typical reactive patterns look like.

4.7 Interpreting Divergence

These examples illustrate that divergence between synthetic and live estimates should not be uniformly interpreted as model failure. In some cases, divergence may reflect:

- Reduction of known measurement errors (satisficing, social desirability)
- Representation of guideline-adherent rather than typical practice
- Smoothing of sampling noise while preserving substantive patterns

The key is understanding *why* estimates diverge and whether the divergence provides analytical value. Questions showing higher KL divergence (Q9, Q15) or moderate RBO (Q9) still preserve clinically meaningful patterns and may better represent certain research objectives than noisy live samples.

5 Implications and Best Practices

5.1 When to Trust Distributional Divergence

Not all divergence indicates model failure. Synthetic estimates may diverge from live samples while still providing value when:

- Divergence may reflect reduction of known measurement errors (satisficing, social desirability bias, central tendency)
- Ranking similarity remains strong (high RBO for multi-select questions)
- Differences may align with clinical guidelines or evidence-based best practices
- Live sample shows patterns consistent with respondent fatigue or cognitive shortcuts
- The research objective is understanding optimal rather than typical practice patterns

5.2 Complementary Use Cases

Synthetic data is most valuable when used to:

- Identify top priorities and dominant patterns (validated by strong RBO performance)
- Reduce noise in exploratory research where exact percentages matter less than relative rankings
- Simulate guideline-adherent practice for benchmarking or protocol development
- Generate hypotheses for validation with targeted live samples
- Understand what engaged practitioners do when following best practices
- Smooth sampling variability in small or moderate-sized studies

5.3 When Live Data Remains Essential

Live surveys should be preferred for:

- Regulatory submissions requiring documented human responses
- Measuring actual (vs. ideal) practice adoption rates for market sizing
- High-stakes decisions based on specific numeric thresholds
- Contexts where social desirability bias is the signal of interest
- Studies where variance and uncertainty quantification are critical

5.4 Evaluation Principles

When evaluating synthetic survey data:

- Match metrics to question types (distributional for single-select, ranking for multi-select)
- Interpret divergence in context of known measurement error patterns
- Complement quantitative metrics with substantive expert review
- Consider whether research objectives align with "typical" vs. "optimal" practice measurement
- Report both successful validation and informative divergence transparently

6 Conclusions

Synthetic survey estimates can provide stable and internally consistent approximations of opinion structure when appropriately constrained and evaluated. The objective is not perfect replication of a single live sample, but preservation of meaningful patterns with reduced noise.

This validation study demonstrates that synthetic physician survey data achieves good to excellent agreement on single-select questions (55% with KL ≤ 0.15) and strong ranking agreement on multi-select questions (90% with RBO ≥ 0.70). Importantly, instances where synthetic estimates diverge from live samples may reflect reduction of measurement error or representation of guideline-adherent practice rather than model failure.

When used responsibly and with appropriate evaluation frameworks, synthetic approaches can complement traditional survey methods and support faster, more robust insight generation. The key is understanding what synthetic data measures—often a less-noisy estimate of opinion structure rather than a literal reproduction of a specific live sample's idiosyncrasies.

A Detailed Validation Results

A.1 Single-Select Questions - KL Divergence

Table 4: Single-Select Questions - KL Divergence

Question	Topic	KL Div.	Agreement
Q5	Familiarity with "sarcopenia"	0.044	Excellent
Q10	Screening time required	0.080	Excellent
Q14	Screening frequency for at-risk patients	0.066	Excellent
Q22	Patients for whom diet/exercise sufficient	0.138	Excellent
Q28	Most helpful educational tools	0.106	Excellent
Q30	Years in practice	0.142	Excellent
Q33	Patient location (urban/rural/suburban)	0.027	Excellent
Q34	Patient gender distribution	0.009	Excellent
Q36	Patient income levels	0.091	Excellent
Q37	Patient living situation	0.068	Excellent
Q16b	Characteristic: Low physical activity	0.081	Excellent
Q16d	Characteristic: Unhealthy diet	0.057	Excellent
Q16e	Characteristic: Social isolation	0.111	Excellent
Q16g	Characteristic: Loss of spouse	0.140	Excellent
Q16h	Characteristic: Move to long-term care	0.134	Excellent
Q23	Referral frequency	0.112	Excellent
Q6	Estimated prevalence of sarcopenia	0.273	Good
Q7	Terminology used in charts	0.242	Good
Q8	Who expresses initial concern	0.298	Good
Q15	Follow-up assessment frequency	0.208	Good
Q16a	Characteristic: Depressive symptoms	0.166	Good
Q16c	Characteristic: Low income	0.165	Good
Q16f	Characteristic: Former high-activity	0.257	Good
Q16i	Characteristic: Recent hospitalization	0.287	Good
Q31	% of 65+ patients in long-term care	0.222	Good
Q35	Patient military status	0.207	Good
Q19	Diagnostic criteria used	0.527	Moderate
Q32	Advanced geriatric training	0.662	High div.*

*Q32 divergence reflects specialty mix (geriatric specialists have higher training rates) rather than model error.

A.2 Multi-Select Questions - RBO Performance

Table 5: Multi-Select Questions - RBO Performance

Question	Topic	RBO	Agreement
Q26	Reasons patients fail to address condition	1.000	Perfect
Q20	Most common treatment recommendations	0.900	Excellent
Q29	Sources of authoritative information	0.890	Excellent
Q24	Primary motivations to screen/treat	0.882	Excellent
Q11	Life events that prompt screening	0.981	Excellent
Q13	Diagnoses that prompt screening	0.809	Excellent
Q18	Measurements to confirm diagnosis	0.773	Excellent
Q25	What would encourage more screening	0.761	Excellent
Q17	ICD-10 codes used for diagnosis	0.748	Excellent
Q9	Screening measures/tools used	0.673	Good

Note: All multi-select questions achieved good to excellent ranking agreement, with RBO values ranging from 0.673 to 1.000. Top-ranked items showed strongest agreement across all questions.

A.3 Questions With Notable Divergence - Clinical Interpretation

Table 6: Questions With Notable Divergence - Clinical Interpretation

Question	Divergence Type	Clinical Interpretation
Q9	Higher evidence-based tool adoption	Synthetic may reflect guideline-following practice; live reflects actual adoption lag
Q8	Higher physician-initiated screening	Synthetic may reflect proactive clinical engagement vs. reactive practice
Q15	More distributed responses	Synthetic may reduce social desirability bias toward "every visit" response
Q16f	Less central tendency	Synthetic may reduce satisficing toward "sometimes" middle category
Q32	Higher training rates	Reflects specialty composition (16% geriatricians + 30% PM&R) rather than error

References

[1] Asch, D. A., Jedrziewski, M. K., & Christakis, N. A. (1997). Response rates to mail surveys published in medical journals. *Journal of Clinical Epidemiology, 50*(10), 1129–1136.

[2] Biemer, P. P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly, 74*(5), 817–848.

[3] Cabana, M. D., Rand, C. S., Powe, N. R., Wu, A. W., Wilson, M. H., Abboud, P. A., & Rubin, H. R. (1999). Why don't physicians follow clinical practice guidelines? A framework for improvement. *JAMA, 282*(15), 1458–1465.

[4] Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). Wiley.

[5] Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American, 236*(5), 119–127.

[6] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). CRC Press.

[7] Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology, 23*(2), 127–135.

[8] Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology* (2nd ed.). Wiley.

[9] Groves, R. M., & Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly, 74*(5), 849–879.

[10] James, W., & Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 361–379.

[11] Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213–236.

[12] Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537–567.

[13] Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*(1), 79–86.

[14] Lohr, S. L. (2019). *Sampling: Design and Analysis* (2nd ed.). Chapman & Hall/CRC.

[15] Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis, 12*(4), 375–385.

[16] Si, Y., Reiter, J. P., & Hillygus, D. S. (2015). Bayesian latent pattern mixture models for handling attrition in panel studies with refreshment samples. *Annals of Applied Statistics, 9*(1), 181–207.

[17] Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly, 70*(1), 66–77.

[18] Snoke, J., Raab, G. M., Nowok, B., Dibben, C., & Slavkovic, A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A, 181*(3), 663–688.

[19] Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

[20] Tourangeau, R., Couper, M. P., & Conrad, F. (2007). Color, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly, 71*(1), 91–112.

[21] VanGeest, J. B., Johnson, T. P., & Welch, V. L. (2007). Methodologies for improving response rates in surveys of physicians: A systematic review. *Evaluation & the Health Professions, 30*(4), 303–321.

[22] Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems, 28*(4), Article 20.