Do synthetic physician panels reproduce sub-national variation in prescribing? A pre-registered negative result against Medicare Part D

Ruly Altamirano

Ereace — QualiSynth / StrataSynth, Barcelona, Spain · Correspondence: hello@qualisynth.com

Preprint v1.1 · 2026-09-18

Version history. v1.0 (2026-09-15) reported the result; v1.1 (2026-09-18) adds the AI disclosure statement, the declarations (Section 12) and the reference list, and names the engine under test. No result, figure or table was changed.

Status. The primary hypothesis was **refuted**. The secondary hypothesis, which served as a contamination control, is reported as **non-conclusive**: it never had a power gate and the observed difference is not distinguishable from chance. Section 7 reports an exploratory observation that is not part of the pre-registration and is labelled as such throughout.

Pre-registration. The design was sealed with a SHA-256 digest published alongside it, before any data were generated. The sealed document fixes the hypothesis, the metric, the threshold, the sample size rule, the classification procedure, and a commitment to publish the outcome whatever it was.

Abstract

Large language models are increasingly used to generate *synthetic respondents* for market and health-services research. Claims that such panels reproduce real-world heterogeneity are common; pre-registered tests against external ground truth are not.

We pre-registered — and sealed by SHA-256 before generating any data — a test of whether **our own** commercial synthetic-respondent engine, **StrataSynth**, reproduces the **state-level gradient in SGLT2 inhibitor adoption** among United States primary care physicians, benchmarked against the **CMS Medicare Part D Prescriber Public Use File**. We generated 2,500 synthetic prescriber interviews (50 states × 50 respondents, zero failures).

The hypothesis was refuted. The rank correlation with real adoption was −0.039 (permutation $p = 0.79$); after the pre-declared attenuation correction, −0.163. More informative than the correlation is the dispersion: between-state variation in the synthetic panel (sd = 0.0699) was indistinguishable from pure sampling noise (expected sd = 0.0685 at a mean $n = 42.5$). Against a simulated null of fifty states with **no between-state signal whatsoever**, the observed dispersion falls at the **60th percentile** on the proportion scale and the **73rd** on the logit scale.

We then tested the three obvious alternative explanations, and none survives. **(i) Measurement ceiling** — ruled out by the logit analysis, where no ceiling exists. **(ii) Instrument wording** — ruled out by a second, count-based measure aligned to what the administrative data actually count, which lowered the base rate from 0.83 to 0.68 and left the between-state difference flat. **(iii) Population mismatch** — correct as a diagnosis, but restricting the real denominator by prescriber activity across 242,081 prescribers raises the real rate to

0.676 in the most active quartile, close to the synthetic 0.68, while leaving state rank order (Spearman 0.89–0.97) and the refutation unchanged.

An exploratory observation is reported separately: the same panel named **prior authorization** as the dominant access barrier, while the contemporaneous Part D formulary files show it applied to **0.42%** of plans for this drug class.

We read these results as a single pattern: the engine produces a population whose prescribing *level* is appropriate for the kind of physician requested, yet **does not distinguish one state from another**, and describes sector-level discourse rather than the specific programme it was instructed to describe.

AI disclosure statement. All respondent-level data analysed in this study are synthetic: they were generated by a large language model through our own commercial synthetic-respondent platform, StrataSynth (see the conflict of interest declaration, Section 12.3). Generating them is the object of study, not a shortcut in its conduct. No human being was interviewed, surveyed or observed at any point, and no individual-level patient record was accessed. The only non-synthetic data are aggregate, publicly published administrative files from the Centers for Medicare & Medicaid Services (2026a, 2026b) and the National Library of Medicine's RxNorm (Nelson et al., 2011), used exclusively as an external benchmark.

The confirmatory analysis reported in Sections 4 to 6 uses no language model at any point: adoption timing was classified from the free-text responses by a deterministic, rule-based procedure, so that the measurement instrument could not introduce variance of its own. **The exploratory observation in Section 7 is different, and is declared as such.** The findings and supporting quotations reported there were extracted from the transcripts by a language model. The safeguard is therefore not the absence of a model but the redundancy imposed on it: three independent extraction runs, only findings present in all three retained (47 of 47), and every supporting quotation verified verbatim against its source transcript (129 of 129).

Generative AI tools were additionally used to assist with drafting and copy-editing of this manuscript. The author reviewed, verified and edited all resulting text and takes full responsibility for the content.

Keywords: synthetic respondents · large language models · simulated agents · external validity · pre-registration · Medicare Part D · prescribing behaviour · health services research

1. Why this question

Synthetic respondents are attractive precisely where real ones are expensive and slow: physicians, specialists, low-incidence populations. Claims that language models can stand in for human samples, and reproduce their heterogeneity, are by now common (Argyle et al., 2023; Aher et al., 2023; Bail, 2024), and the practice has spread into applied research and commercial user research (Kapania et al., 2025; Lewis & Sauro, 2026a, 2026b). The critical question for any applied use is not whether the transcripts *read* plausibly — they do — but whether the resulting distributions carry information about the world.

The evidence is genuinely mixed, and it is worth stating in both directions. On the favourable side, Ashokkumar et al. (2026) find that language models predict the results of social science experiments with substantial accuracy, though the authors themselves report that effect sizes are overestimated. On the unfavourable side, a consistent finding is *flattening*: diminished diversity of thought relative to human samples (Park et al., 2024), misportrayal and homogenisation of identity groups (Wang et al., 2025), and the more general caution that models are not interchangeable with human participants (Dillion et al., 2023). Within health specifically, the record cuts both ways, with synthetic cohorts falling short of epidemiological validity in at least one published assessment (Fuentes-Martín et al., 2026).

Most published evaluations compare synthetic answers against survey benchmarks (Santurkar et al., 2023; Bisbee et al., 2024; Boelaert et al., 2025; Shrestha et al., 2024), which share the self-report problem with the thing being evaluated. We chose a benchmark that does not: **administrative prescribing records**.

Within that, we chose a cut that is not part of the summary literature a model is most likely to have absorbed: the geographic gradient in adoption **within a single specialty**, which is a by-product of local payer policy and panel composition rather than a headline finding.

We should be explicit that this is a weaker claim than "there is nothing to recall". Sub-national variation in adoption of this drug class *is* documented in the specialised literature — at state level among cardiologists (Kim et al., 2026), across the 306 Dartmouth hospital referral regions (Chen et al., 2024), and at county level, where a median 6.2% of metformin-treated beneficiaries received an SGLT2 inhibitor with an interquartile range of 3.4% to 9.2% across 3,066 counties (Hanna et al., 2022). Recall therefore cannot be excluded a priori, and we do not exclude it. What settles the matter here is the direction of the result rather than the design: **an engine that produced no between-state gradient at all cannot have recalled one.** Had the gradient appeared, distinguishing simulation from retrieval would have required a further study; it did not appear, and that question does not arise.

2. Pre-registration and entry conditions

Three entry conditions had to pass before generation was permitted, and all three did.

Gate	Criterion	Result
Calibration	Reproduce an already-published figure from the raw file, end to end	Pass
Power gate	Smallest n per state reaching power ≥ 0.80 for $\rho_v = 0.6$ at the measured real dispersion	sd = 0.0548 → **n = 50 per state**, power 0.815
Clustering pilot	ICC below 0.02, else re-declare the effective n	150 interviews, **ICC estimate truncated at zero** (see note)

Note on the clustering pilot. The ANOVA estimator returned a negative value and was truncated at zero, because within-state variance exceeded between-state variance. With three groups this is an imprecise estimate: it does not rule out a small true ICC, and the defensible reading is *no clustering detected* rather than *no clustering*. Either way it does not increase the required sample size, which is the decision this gate governed.

H1 (primary). Within primary care, the synthetic population reproduces the state ordering of adoption intensity observed in Part D. Metric: Spearman's ρ. Threshold: **$\rho \geq 0.40$ with permutation $p < 0.05$**. Significance by permutation rather than the asymptotic approximation, because ties are expected.

Attenuation correction was declared in advance, not after the fact. The measurement error here is known and exact: binomial with declared n on the synthetic side, against a census on the real side. The sealed rule is *disattenuated ρ as the primary analysis, crude ρ as secondary, both always published, with the estimated reliability alongside*. Declared before, this is method; declared afterwards, it would be cosmetics.

H2 (secondary, declared weak in advance). The ordering of adoption across the four specialties matches the published literature on prescribing patterns in this class among Medicare beneficiaries (Sangha et al., 2021). Expected to hold; its function was contamination control, never evidence.

3. Data and construction

3.1 Ground truth

CMS Medicare Part D Prescriber Public Use File (Centers for Medicare & Medicaid Services, 2026a), data year 2021, two datasets joined by NPI: *by Provider and Drug* for the numerator and *by Provider* for the denominator. Primary care was defined as Family Practice, Internal Medicine and General Practice; SGLT2 monotherapy products only; physicians only, excluding nurse practitioners and physician assistants. The District of Columbia and the territories were excluded *before sealing*: they were the three lowest rates and inflated the dispersion by 10.7%.

3.2 Synthetic panel

Fifty segments, one per state, each declaring the state, the profession and a predominantly-Medicare practice. `demographics.profession` was set explicitly on every biography. This is not an implementation detail: prior internal measurement found that prose description alone does not fix occupation — six of six personas were off-profile without the field, zero of six with it.

3.3 Anti-tautology clause

Biographies were forbidden, in any form including implication, from containing adoption-timing language, any attitude toward the drug class or toward payers, and any prescribing-behaviour claim regarding the class. No psychometric seed was fixed. The panel of fifty segments and 2,500 personas was constructed through the public API of StrataSynth, the engine under test (see the conflict of interest declaration, Section 12.3).

3.4 Classification

Adoption timing was extracted from free text by a **deterministic** procedure — explicit year first, then a declared list of qualitative expressions — with no language-model judge. The reason is that a judge introduces its own variance into the quantity being measured. The cost of this choice is that responses fitting no rule are reported as unclassified rather than forced into a category.

4. Result

2,500 of 2,500 interviews completed, zero failures. 2,123 classified (85%), 377 unclassifiable, and **zero without a region**.

Measure	Value	Role
Spearman ρ, crude	−0.0391	secondary
Permutation *p* (10,000 label permutations)	0.7888	significance
Reliability of the measure	0.0577	input to the correction
Spearman ρ, disattenuated	**−0.1626**	**primary**
Sealed threshold	≥ 0.40	a priori

Both figures fall below zero. **H1 is refuted.** There is no tension between the crude and corrected values: the disattenuation rescues nothing because there is nothing to rescue.

4.1 The dispersion is the informative quantity

	Synthetic	Real (CMS)
Mean	0.725	0.259
sd between states	**0.0699**	**0.0548**
Range	0.610 – 0.867	0.166 – 0.371

At first sight the synthetic dispersion appears *larger* than the real one. It is not informative: with a mean of 42.5 classified responses per state at $p = 0.725$, pure binomial noise already yields sd = $\sqrt{(p(1-p)/n)}$ = **0.0685**.

Rather than rely on an analytic variance decomposition — a point estimate that can be disputed on scale grounds — we simulated the null directly: **4,000 replications of fifty states with no between-state signal at all**, at the observed per-state n and the observed synthetic mean.

Scale	Observed sd	Null sd	Percentile of the observation
Proportion	0.0699	0.0683	**60**
Logit	0.3821	0.3587	**73**

> The observed between-state dispersion sits at the 60th percentile of a world containing no signal. The engine did not generate fifty populations; it generated one population fifty times.

This also closes a gap a careful reader will find in the design. The power gate was set at $\rho_v = 0.6$ while the decision threshold was $\rho \geq 0.40$, leaving a band in which a true correlation could exist and the study would be underpowered to detect it. The dispersion analysis answers that directly: with between-state variance indistinguishable from zero, **there is no signal of any strength available to correlate**. The refutation does not rest on the rank correlation's power alone.

The direction of the classification noise matters here and works in favour of the conclusion. An imperfect classifier adds variance *within* each state, never between states, so it can only inflate the observed dispersion. The residual signal is therefore an upper bound, not an estimate.

5. Robustness: three alternative explanations, tested

5.1 Was the measure saturated?

At a base rate of 0.83 in the follow-up arms, a ceiling could in principle compress variation. This is directly testable, because **on the logit scale there is no ceiling**. The observation still sits at the 73rd percentile of the null. Furthermore, at $n \approx 95$ and $p = 0.83$ the standard error of a between-state difference is 0.054, so a difference of 0.15 would have been 2.8 standard errors. There was ample room to detect one.

5.2 Was it the wording?

Our timing question asks when the respondent *first* prescribed, which measures an ever-event; the administrative benchmark counts prescribing *within the year*. These are different constructs, and that is a genuine defect of the instrument independent of the ceiling question. We therefore ran a second arm changing **one question only**, to a recent count — "over the last six months, how many patients have you started" — with an adoption threshold declared in advance.

Both follow-up arms used the **two most widely separated states in the real data** — Alaska (0.1658) and Kentucky (0.3710), a real gap of **0.2052** — with 100 respondents each, and a pass criterion written and version-controlled before the data existed ($d \geq 0.15$ with $p < 0.05$).

Instrument	Alaska	Kentucky	*d*	*p*
No regional context	0.756	0.837	+0.081	n.s.
Regional context, binary measure	0.837	0.821	**−0.016**	0.77
Regional context, count measure	0.697	0.663	**−0.034**	0.61
Real (CMS)	**0.166**	**0.371**	**+0.205**	—

The count measure moved the base rate from 0.83 to 0.68 — off the ceiling — and the difference remained flat, with the sign reversed. **The instrument is not the explanation.**

The regional context supplied in those arms was real and drawn from the same census file — panel size, mean beneficiary age, HCC risk score, dual-eligible share, low-income-subsidy share — and **never the prescribing rate itself**, which would have manufactured the result. Between these two states, panel size differs by a factor of 2.5 and mean clinical risk by 20%. The engine received that and did not change behaviour.

5.3 Were we comparing different populations?

This objection is correct as a diagnosis. The benchmark denominator includes every primary care prescriber in the file, including minimal-panel and part-time prescribers; the synthetic biographies are all active physicians with full panels. Prescriber- and patient-level predictors of adoption in this class are themselves unevenly distributed (Eberly et al., 2021), so the composition of the denominator is not a neutral choice.

Restricting the real denominator by prescriber volume, across **242,081 prescribers**:

Denominator	Real rate	Spearman vs. sealed ordering	H1 recomputed
All prescribers (sealed)	0.2591	1.000	−0.039
Excluding least active quartile	0.3451	0.973	−0.065
Most active half	0.4985	0.948	−0.052
Most active quartile	**0.6763**	0.891	−0.047

The synthetic count measure returned 0.697 and 0.663; the most active quartile of the real data returns 0.6763. **The synthetic level is consistent with the most active quartile.**

> The quartile was identified after observing where the synthetic figure fell. This is an **observed correspondence, not a calibration test**: no quartile correspondence was declared in advance. It suffices to dismiss the "different populations" objection; it does not establish calibration.

Crucially, state rank order survives restriction (Spearman 0.89–0.97 against the sealed ordering), and H1 remains negative under all four denominators.

6. The secondary hypothesis, and why it is reported as unresolved

Specialty	*n*	Readable	Early or prior adoption
Endocrinology	25	21	0.905
Primary care	25	19	0.789
Nephrology	25	21	0.667
Cardiology	25	18	0.556

The single relation the pre-registration committed to — primary care *below* cardiology — is not observed. **But with 19 and 18 readable responses the difference is not distinguishable from chance** ($z = 1.52$, $p = 0.13$), and **H2 never had a power gate**: the pre-registration set one for H1 only.

> H2 is therefore reported as **non-conclusive, not as refuted**. Declaring a negative where the data support only indeterminacy would be the error this paper exists to avoid. The operational definition of early adoption across specialties was also ours, not the pre-registration's.

The consequence is that the outcome anticipated in the publication schedule — *"H1 fails, H2 holds, therefore the model recalls the literature rather than simulating local environments"* — **cannot be invoked**, because its second half was never established.

7. An exploratory observation on friction narratives

A second arm of 100 respondents and 1,200 responses explored administrative friction. Its methodological value is that each finding was extracted **three times independently** and only findings present in all three were retained: 47 of 47 survived, with 129 of 129 supporting quotations verified verbatim against the transcripts.

The substantive content is where it becomes interesting. The panel named **prior authorization** as the leading barrier in both physician groups, and **step therapy** as the most frequently cited response to payer requirements. The biographies declared predominantly Medicare practice. We checked this against the contemporaneous **CMS Prescription Drug Plan Formulary file (2021Q3)** (Centers for Medicare & Medicaid Services, 2026b) — 1.6 million formulary rows, 10,006 of them SGLT2 products identified through 86 RxNorm product concepts (Nelson et al., 2011), across 456 formularies, counted by unique plan.

	As reported by the panel	In the declared programme
Prior authorization	leading barrier	**0.42%** of plans (sd 1.72)
Step therapy	most-cited payer response	**3.55%** of plans (sd 5.20)

This is exploratory, was not pre-registered, and carries three limits that must travel with it.

- **The instrument did not scope it.** The question reads "ran into a payer requirement" and **does not name Medicare**. A respondent describing a commercially-insured patient is answering correctly. Part of the distance may come from the wording rather than from the engine.
- **Part D is not the whole market.** Prior authorization is common in commercial and Medicaid coverage, and formulary restrictions on this class within Part D have themselves changed over time (Luo et al., 2020; Buttorff et al., 2025), which is why the comparison is made against the contemporaneous file rather than against a general characterisation of the programme. The defensible statement is narrow: *the friction described does not correspond to the programme the panel was told it worked in.*

- **The 0.42% is counted per plan.** Weighted by enrollment it is a different figure, and that is the one relevant to real-world exposure. It has not been computed.

8. Limitations

- **One engine, one base model.** Nothing here transfers to other systems without measuring them.
- **One drug class, one country, one year.**
- **Reliability of 0.058.** The deterministic classifier is noisy. With that reliability an attenuation correction amplifies whatever it is given — which is precisely why the primary argument in this paper is the dispersion, which does not require it.
- **15% unclassifiable.** 377 of 2,500 responses fitted no rule. If they are structurally different from the rest, the correlation is biased in a direction this design cannot determine.
- **No human arm**, and none is required here, since the administrative record *is* the real data. The study says nothing about convergence with human interviews.
- **Two alternative instruments and four denominators**, not all possible ones. Strong evidence, not certainty.
- **Development environment.** All generation ran in the platform's test environment.

9. What this licenses, and what it does not

9.1 Licensed

- This engine, asked to produce state-level variation, did not produce it: between-state variation is indistinguishable from sampling noise on both the proportion and the logit scale.
- The declared region **reaches** the persona — zero of 2,500 lacked one. What does not arrive is the behaviour associated with it.
- The synthetic prescribing *level* is consistent with the comparable real population — **an observed correspondence, not a calibration test** (Section 5.3).

9.2 Not licensed

- That synthetic populations in general fail to reproduce geography. One engine was measured.
- That the engine ignores geography altogether. Between-country structure has been measured elsewhere (Altamirano, 2026); what this study constrains is its **origin** — country and language conditioning rather than the persona layer.
- That the gradient is unattainable. The finding is that it is **not obtained by asking for it**, and not obtained by supplying panel-composition context.
- Anything clinical. Efficacy, safety and treatment comparison were not studied.

10. Reproducibility

The pre-registration and its SHA-256 digest, the analysis scripts — each carrying an offline self-test suite that turns red when the defect it guards against is reintroduced — and the raw per-state figures are **available on request today**. A reader who asks can verify the digest against the sealed design and re-run every figure in this paper.

We state this in the present tense deliberately. The argument of this note is that the reason to trust a result is that it can be checked; a promise of future deposit does not support that argument at the moment it is read.

Every figure in this paper derives either from public data (CMS Part D Prescriber Public Use File, CMS Prescription Drug Plan Formulary files, and RxNorm) or from synthetic responses generated through a documented API path.

11. Declaration on the nature of the study

This is a 100% synthetic study. No human being was interviewed. All responses were generated by a large language model. The only real data are the aggregate prescribing and formulary records published by CMS, used exclusively as a benchmark.

It is not a clinical study and contains no clinical guidance. It says nothing about efficacy, safety or comparison between treatments. The subject matter is adoption behaviour and administrative friction.

The pre-registration committed to publishing the outcome whatever it turned out to be. This is that outcome.

12. Declarations

12.1 Pre-registration

The design was sealed with a SHA-256 digest published alongside it, before any data were generated. The sealed document fixes the hypothesis, the metric, the decision threshold, the sample-size rule, the classification procedure, and a commitment to publish the outcome whatever it turned out to be. The primary hypothesis was refuted and is reported as such. The secondary hypothesis is reported as non-conclusive rather than refuted, because it never carried a power gate. Section 7 reports an exploratory observation that was not part of the pre-registration and is labelled as such throughout.

12.2 Ethics and human participants

This study involves no human participants. No person was interviewed, surveyed, observed or recruited; no informed consent was required because there was no one to consent; and no individual-level patient data of any kind were accessed. The synthetic respondents are model outputs, not people, and are described as such throughout. The only real-world data are aggregate, publicly available administrative files published by the Centers for Medicare & Medicaid Services, which contain no individually identifiable health information. For these reasons no institutional review board approval was sought or required.

This is not a clinical study. It reports nothing about the efficacy, safety or comparative effectiveness of any medicine, and contains no clinical guidance. Its subject matter is prescribing-adoption behaviour and administrative friction, and its object of measurement is a software system, not a patient population.

12.3 Conflict of interest

The author is the founder of Ereace, the company that develops StrataSynth, the synthetic-respondent engine evaluated in this study, and therefore has a direct financial interest in the system under test. This relationship is material and conditions the reading of the entire article, so it is declared here explicitly rather than in a footnote.

It is declared, moreover, with the detail that allows it to be judged. The study was pre-registered predicting that the engine *would* reproduce the state-level gradient — the product-favourable prediction — and the design was sealed by SHA-256 digest before any data were generated. The data refuted it. The central result of this article

is, in commercial terms, unfavourable to the author: the engine generated one population fifty times rather than fifty populations. The commitment to publish the outcome whatever it turned out to be was registered in writing before any data existed, and this is that outcome.

The limit of the declaration should also be stated: transparency does not remove the conflict. The decisions taken before sealing — which comparison to run, which alternative explanations to pre-commit to testing — were taken by an interested party. What a sceptical reader can verify independently is that no hypothesis was added, removed or reformulated after the data were seen: the sealed pre-registration and its digest allow exactly that check.

12.4 Funding

This work received no external funding. Generation and compute costs were borne by Ereace, the author's company. There was no grant, fellowship or third-party funding, and no funder had any role in the design, conduct, analysis, or decision to publish.

12.5 Data and code availability

The sealed pre-registration and its SHA-256 digest, the per-state figures, the deterministic classifier, and the analysis scripts with fixed random seeds are available from the author on request today, as stated in Section 10: a reader who asks can verify the digest against the sealed design and re-run every figure in this paper. The benchmark data are public and are cited in the references.

12.6 Author contributions (CRediT taxonomy)

Single authorship. Ruly Altamirano is responsible for conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing (original draft, review and editing), and visualization.

References

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, PMLR 202, 337-371.

Altamirano, R. (2026). Procedural decision architectures under uncertainty in synthetic populations: what produces between-market structure, and what does not. *SSRN Working Paper*. https://doi.org/10.2139/ssrn.7176420

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337-351. https://doi.org/10.1017/pan.2023.2

Ashokkumar, A., Hewitt, L., Ghezae, I., & Willer, R. (2026). Large language models can predict the results of social science experiments. *Nature*, 656(8126), 115-122. https://doi.org/10.1038/s41586-026-10742-x

Bail, C. A. (2024). Can generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21), e2314021121. https://doi.org/10.1073/pnas.2314021121

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4), 401-416. https://doi.org/10.1017/pan.2024.5

Boelaert, J., Coavoux, S., Ollion, E., Petev, I., & Prag, P. (2025). Machine bias: How do generative language models answer opinion polls? *Sociological Methods & Research*, 54(3), 1156-1196.

https://doi.org/10.1177/00491241251330582

Buttorff, C., Khodyakov, D., Taylor, E. A., Reid, R. O., Sorbero, M. E., & Dworsky, M. (2025). Trends in Medicare Part D formulary coverage for non-insulin diabetes medications, 2020-2024. *Journal of General Internal Medicine*, 40(7), 1669-1671. https://doi.org/10.1007/s11606-024-09171-1

Centers for Medicare & Medicaid Services. (2026a). *Medicare Part D Prescribers - by Provider and Drug* [Data set]. U.S. Department of Health and Human Services. https://data.cms.gov/provider-summary-by-type-of-service/medicare-part-d-prescribers/medicare-part-d-prescribers-by-provider-and-drug

Centers for Medicare & Medicaid Services. (2026b). *Quarterly Prescription Drug Plan Formulary, Pharmacy Network, and Pricing Information* [Data set]. U.S. Department of Health and Human Services. https://data.cms.gov/provider-summary-by-type-of-service/medicare-part-d-prescribers/quarterly-prescription-drug-plan-formulary-pharmacy-network-and-pricing-information

Chen, W.-H., Li, Y., Yang, L., Allen, J. M., Shao, H., Donahoo, W. T., Billelo, L., Hu, X., Shenkman, E. A., Bian, J., Smith, S. M., & Guo, J. (2024). Geographic variation and racial disparities in adoption of newer glucose-lowering drugs with cardiovascular benefits among US Medicare beneficiaries with type 2 diabetes. *PLOS ONE*, 19(1), e0297208. https://doi.org/10.1371/journal.pone.0297208

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7), 597-600. https://doi.org/10.1016/j.tics.2023.04.008

Eberly, L. A., Yang, L., Eneanya, N. D., Essien, U. R., Julien, H., Nathan, A. S., Khatana, S. A. M., Dayoub, E. J., Fanaroff, A. C., Giri, J., Groeneveld, P. W., & Adusumalli, S. (2021). Association of race/ethnicity, gender, and socioeconomic status with sodium-glucose cotransporter 2 inhibitor use among patients with diabetes in the US. *JAMA Network Open*, 4(4), e216139. https://doi.org/10.1001/jamanetworkopen.2021.6139

Fuentes-Martín, A., Mayol, J., Segura Mendez, B., & Cilleruelo-Ramos, A. (2026). Synthetic lung-cancer cohorts generated by a large language model: Epidemiological validity assessment. *Open Respiratory Archives*, 8(1), 100533. https://doi.org/10.1016/j.opresp.2025.100533

Hanna, J., Nargesi, A. A., Essien, U. R., Sangha, V., Lin, Z., Krumholz, H. M., & Khera, R. (2022). County-level variation in cardioprotective antihyperglycemic prescribing among Medicare beneficiaries. *American Journal of Preventive Cardiology*, 11, 100370. https://doi.org/10.1016/j.ajpc.2022.100370

Kapania, S., Agnew, W., Eslami, M., Heidari, H., & Fox, S. E. (2025). Simulacrum of stories: Examining large language models as qualitative research participants. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*, 1-17. https://doi.org/10.1145/3706598.3713220

Kim, T., Ryan, D., & Joseph, J. (2026). Adoption and diffusion of SGLT2 inhibitors among cardiologists in Medicare Part D, 2019-2023: A provider-level observational study. *BMC Cardiovascular Disorders*. Advance online publication. https://doi.org/10.1186/s12872-026-06460-x

Lewis, J. R., & Sauro, J. (2026a, April 14). *A review of experiments with synthetic users*. MeasuringU. https://measuringu.com/review-of-experiments-with-synthetic-users/

Lewis, J. R., & Sauro, J. (2026b, June 23). *What are the different types of synthetic users?* MeasuringU. https://measuringu.com/what-are-the-different-types-of-synthetic-users/

Luo, J., Feldman, R., Rothenberger, S. D., Hernandez, I., & Gellad, W. F. (2020). Coverage, formulary restrictions, and out-of-pocket costs for sodium-glucose cotransporter 2 inhibitors and glucagon-like peptide 1 receptor agonists in the Medicare Part D program. *JAMA Network Open*, 3(10), e2020969. https://doi.org/10.1001/jamanetworkopen.2020.20969

Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T., & Moore, R. (2011). Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association*, 18(4), 441-448. https://doi.org/10.1136/amiajnl-2011-000116

Park, P. S., Schoenegger, P., & Zhu, C. (2024). Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 56(6), 5754-5770. https://doi.org/10.3758/s13428-023-02307-x

Sangha, V., Lipska, K. J., Lin, Z., Inzucchi, S. E., McGuire, D. K., Krumholz, H. M., & Khera, R. (2021). Patterns of prescribing sodium-glucose cotransporter-2 inhibitors for Medicare beneficiaries in the United States. *Circulation: Cardiovascular Quality and Outcomes*, 14(12), e008381. https://doi.org/10.1161/CIRCOUTCOMES.121.008381

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, PMLR 202, 29971-30004.

Shrestha, P., Krpan, D., Koaik, F., Schnider, R., Sayess, D., & Binbaz, M. S. (2024). Beyond WEIRD: Can synthetic survey participants substitute for humans in global policy research? *Behavioral Science & Policy*, 10(2), 26-45. https://doi.org/10.1177/23794607241311793

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3), 400-411. https://doi.org/10.1038/s42256-025-00986-z