The Patient Digital Twin

AI-Augmented Patient Insights
from Public Health Data

Published by **Simsurveys**

February 2026

Executive Summary

Patient research is among the most valuable and most difficult work in market research. Pharmaceutical companies, biotech firms, health systems, and medical device manufacturers depend on patient insights to guide product development, shape clinical messaging, and understand the real-world experience of living with a condition. But recruiting patients for survey research is slow, expensive, and constrained by privacy regulations that do not apply to consumer or physician panels.

The Patient Digital Twin is a new approach, developed by Simsurveys, that fundamentally changes how patient insights are generated. Instead of recruiting patients through panels, the system builds digital twins from publicly available, HIPAA-compliant federal health datasets—real patient records that have already been de-identified at the source by the U.S. government. Combined with a purpose-built patient AI model, these twins can generate validated survey responses for any patient population, including rare and underserved conditions, in minutes rather than months.

This paper explains why the current approach to patient research is unsustainable, how the Patient Digital Twin works, and what the early validation evidence shows.

The Problem: Patient Research Is Uniquely Constrained

Consumer panels are plentiful. Physician panels are expensive but established. Patient panels barely exist—and for good reason.

Recruitment Is Expensive and Slow

Finding patients with a specific condition who are willing and able to participate in research is fundamentally harder than finding consumers or physicians. Incidence rates for many conditions are low. Patients must be verified as having the condition, must be in a suitable stage of their care journey, and must consent to participation. For a typical patient survey, recruitment alone can take 8–12 weeks and cost thousands of dollars per complete.

Privacy Creates Structural Barriers

Patient data is governed by HIPAA and related regulations that create legitimate but substantial barriers to research. Clinical encounter data—the richest source of patient experience information—sits behind institutional walls. Accessing it requires IRB approvals, data use agreements, and de-identification protocols that can take months to navigate. Many studies never get access at all.

Rare Conditions Are Nearly Impossible

For conditions affecting fewer than 1 in 1,000 people, traditional survey recruitment becomes impractical. The cost per complete can reach hundreds or thousands of dollars. Sample sizes are often too small for statistical reliability. Many rare disease patient studies simply do not happen—not because the insights are not needed, but because the logistics are prohibitive.

The Data Exists—It Is Just Inaccessible

Here is the paradox: the U.S. government surveys hundreds of thousands of patients every year through federally mandated programs. These surveys capture exactly the information that patient researchers need—conditions, treatments, healthcare experiences, functional status, mental health, insurance, and social determinants. The data is de-identified, publicly available, and updated annually.

Yet this data has never been used to build patient digital twins. Until now.

The Solution: Real Patients, Public Data, Purpose-Built AI

The Patient Digital Twin addresses these constraints through a fundamentally different architecture than consumer or physician twins. There is no panel to recruit. No profiling instrument to administer. No incentive to pay. The patients have already been profiled—by the federal government.

Part 1: Federal Health Data as Twin Profiles

The U.S. government conducts several large-scale patient surveys annually, each capturing hundreds of data points per respondent:

- **National Health Interview Survey (NHIS)**—approximately 30,000 households per year, covering chronic conditions, healthcare access, functional status, pain, disability, mental health, and social determinants. Conducted by the CDC.

- **Medical Expenditure Panel Survey (MEPS)**—approximately 18,000 respondents per year, covering healthcare utilization, patient experience, provider quality ratings, expenditures, and insurance. Conducted by AHRQ.
- **National Health and Nutrition Examination Survey (NHANES)**—approximately 15,000 respondents per cycle, combining health interviews with clinical examinations, laboratory measurements, and detailed condition assessments. Conducted by the CDC.

Each respondent in these surveys is a real patient with 120 to 400 structured data points—demographics, diagnosed conditions, current treatments, healthcare experiences, functional limitations, and behavioral patterns. All data is de-identified at the source through rigorous federal disclosure avoidance protocols. No protected health information is present.

These records become the seed profiles for patient digital twins. Each twin is one real person from one dataset—a complete, internally consistent profile of how that patient lives with their condition and interacts with the healthcare system.

Part 2: A Purpose-Built Patient AI Model

The seed profiles provide the individual foundation. The AI model provides the inference capability.

Simsurveys' patient model is trained on the full breadth of these federal datasets—hundreds of thousands of real patient records spanning chronic conditions, acute care experiences, mental health, disability, and preventive care. The model learns the patterns that connect a patient's condition, demographics, treatment history, and healthcare experiences to how they respond to research questions.

When a new patient study is fielded, the system identifies relevant twins based on condition, demographics, and other targeting criteria, then generates responses that are consistent with each twin's established profile. The result is a complete patient dataset—delivered in minutes, grounded in real patient data, and validated against national benchmarks.

> This is not synthetic data generated from population averages. Each response is anchored to the profile of a specific real patient whose data was collected through a rigorous federal survey program.

Condition Targeting: The Core Capability

What makes patient digital twins fundamentally different from consumer or physician twins is **condition-based targeting**. The federal datasets contain diagnosed conditions for every respondent, enabling precise selection of twins by disease state.

Common Conditions

For prevalent conditions—diabetes, heart disease, asthma, COPD, hypertension, arthritis, depression, cancer—the twin database contains thousands of real patient profiles. Studies targeting these populations can be fielded immediately with high statistical confidence.

Rare and Underserved Conditions

For less common conditions, the combined dataset of over 500,000 respondents provides meaningful sample sizes even at low prevalence rates. A condition affecting 0.1% of the population yields approximately 500 real patient profiles across the combined datasets.

For conditions with very small populations, Simsurveys has developed methods to generate additional patient profiles that are statistically consistent with the real patients in the dataset. These methods ensure that even rare disease studies can achieve viable sample sizes while remaining anchored to real patient data.

The key advantage: Patient research that was previously impossible due to recruitment constraints—rare diseases, specific treatment stages, underserved populations—becomes feasible through condition-targeted digital twins.

No Panel Partner Required

Consumer and physician digital twins require a panel partner—an organization that recruits real people to complete profiling instruments. This creates dependencies on recruitment pipelines, incentive budgets, and ongoing panelist management.

The Patient Digital Twin eliminates this dependency entirely:

	Consumer / Physician Twin	Patient Twin
Seed data source	Panel recruitment	Federal public-use data
Acquisition cost	$50–200 per recruit	Zero
Profiling instrument	100–250 questions	Already completed (120–400 data points)
Available profiles	Limited by recruitment	500,000+ and growing
Annual refresh	Re-profile panelists	New government data released annually
HIPAA concern	Minimal	Solved at source
Fraud risk	Panel fraud issues	Federal survey quality controls

The economics are structurally different. There is no per-twin acquisition cost, no incentive budget, no panel attrition to manage. The U.S. government funds the data collection. Simsurveys builds the twins and the model.

Validation: Early Evidence

Simsurveys conducted a baseline validation of the patient model against the HCAHPS (Hospital Consumer Assessment of Healthcare Providers and Systems) survey—the CMS-mandated national standard for measuring patient experience across U.S. hospitals. HCAHPS is publicly reported, covers 21 questions across seven domains, and represents approximately 631,000 completed surveys from 4,304 hospitals.

Study Design

Simsurveys generated 1,000 simulated patient respondents using the HCAHPS questionnaire with embedded skip logic. Results were compared against weighted national averages computed from hospital-level CMS data (January–December 2024 discharges, October 2025 public report).

Results

Domain	Avg. KL Divergence	Assessment
Nurse Communication (Q1–Q3)	0.072	Excellent
Doctor Communication (Q4–Q6)	0.068	Excellent
Hospital Environment (Q7–Q8)	0.092	Excellent
Staff Responsiveness (Q9, Q11)	0.110	Strong
Medicine Communication (Q13–Q14)	0.058	Excellent
Care Transition (Q17–Q19)	0.048	Excellent
Overall Rating & Recommend (Q20–Q21)	0.065	Excellent

All 19 measured questions achieved KL divergence below the 0.15 threshold for strong distributional alignment. The majority fell below 0.08, indicating excellent agreement.

Significance

This validation was conducted using the **general consumer model**—before any patient-specific fine-tuning. The model had never been trained on inpatient experience data, which is governed by HIPAA and was not present in the training corpus. Despite this, the model accurately reproduced the distributional patterns of real hospital patient experience across all seven HCAHPS domains.

Fine-tuning on the patient-specific federal datasets (NHIS, MEPS, NHANES) is expected to substantially improve these already-strong results, particularly for condition-specific studies.

The Flywheel

Unlike traditional patient research infrastructure, which must be rebuilt for each study, the Patient Digital Twin system improves automatically over time:

The data grows. NHIS, MEPS, and NHANES release new data annually. Each release adds tens of thousands of new patient profiles to the twin database—without any recruitment effort or cost.

The model improves. Every new dataset incorporated into training makes the patient model more accurate across conditions, demographics, and healthcare contexts.

Rare conditions become feasible. As the cumulative twin database grows year over year, the number of real patients available for rare condition studies increases steadily. Conditions that required synthetic supplementation in year one may have sufficient real patient profiles by year three.

The key insight: The U.S. government is effectively funding the ongoing expansion of the patient twin database through its annual survey programs. Simsurveys converts that public investment into a commercially viable patient research infrastructure.

Applications

The Patient Digital Twin enables patient research that is currently impractical or impossible:

- **Pharmaceutical market research**—patient experience studies, treatment satisfaction, burden of illness, unmet needs assessment, and competitive landscape analysis across any condition
- **Rare disease insights**—patient-reported outcomes and experience data for conditions where traditional recruitment cannot achieve viable sample sizes
- **Health system quality measurement**—rapid benchmarking against national patient experience standards without fielding new surveys
- **Medical device research**—patient experience with devices, procedures, and post-surgical outcomes across targeted condition populations
- **Health equity research**—oversampling underrepresented patient populations that are present in federal datasets but difficult to recruit through traditional panels
- **Regulatory support**—patient preference and experience data to support FDA submissions, HTA evaluations, and value dossiers

Conclusion

Patient research has been constrained by a fundamental mismatch: the demand for patient insights is growing, but the infrastructure for generating them has not kept pace. Recruitment is slow. Privacy requirements are complex. Rare conditions are underserved. And the richest patient data in the country—collected annually by the federal government—has never been leveraged for commercial research.

The Patient Digital Twin closes this gap. By building digital twins from real, de-identified patient records and combining them with a purpose-built patient AI model, Simsurveys delivers patient insights at a speed, cost, and scale that traditional methods cannot match. Early validation against national HCAHPS benchmarks confirms that the approach produces research-grade results—and this is before patient-specific model fine-tuning.

The data exists. The model works. The validation is underway.

About Simsurveys

Simsurveys operates the leading AI-augmented survey research platform, with domain-specific models for consumer, healthcare professional, and patient research. The platform supports digital twin creation, synthetic data generation, augmented data extension, and real-time preference inference through the Oracle API. Simsurveys' patient model is trained on federally collected patient data and validated against national healthcare benchmarks. The methodology is protected by pending U.S. patents.

For inquiries about the Patient Digital Twin, contact **info@simsurveys.com** or visit **simsurveys.com**.