The CMS national provider directory: a complete analysis of 27.2 million healthcare records in the entrepreneurial opportunity that they represent

APR 18, 2026

I. Introduction: The Infrastructure That Was Missing

For decades, the United States healthcare system has operated without a single authoritative, machine-readable directory of its providers. Hospitals, insurers, health systems, and technology companies each maintained their own proprietary provider databases - expensive to build, difficult to maintain, and impossible to reconcile with one another. A physician might appear in dozens of databases simultaneously, each with slightly different information about their specialty, location, affiliations, and contact details. This fragmentation imposed enormous costs on the system: prior authorization delays, misdirected referrals, failed care coordination, and billions of dollars spent annually on provider data management by organizations that would rather spend that money on care.

On April 9, 2026, the Centers for Medicare and Medicaid Services (CMS) released the National Provider Directory (NPD) - a single, public, FHIR-formatted dataset containing every Medicare-enrolled provider in the United States. The release, available at https://directory.cms.gov is the most comprehensive public healthcare provider dataset ever assembled. It contains 27,204,567 records across six FHIR resource types, compressed to 2.8 gigabytes and freely downloadable by anyone.

This essay presents the results of a complete analysis of every record in the dataset - not a sample, not an approximation, but a full population analysis of all 27.2 million records. The analysis was conducted using Python streaming scripts, with the most computationally intensive cross-resource graph linkage analysis run on GitHub Actions cloud infrastructure to avoid local compute constraints. The findings reveal both the extraordinary power of what CMS has released and the significant gaps that remain - gaps that represent direct entrepreneurial opportunities for health technology builders.

Alongside this analysis, a working prototype was built to make the data tangible and interactive: the CMS NPD Explorer, available at onhealthcare.manus.space. The application is a six-page React 19 web application built with TypeScript, Tailwind CSS 4, and Recharts, deployed on Manus cloud infrastructure. It was designed with a Federal Data Observatory aesthetic - a deep navy sidebar, Source Serif 4 display typography paired with DM Sans for body text, and a dark-on-light color system that evokes institutional precision rather than consumer-product softness. The site includes an Overview Dashboard displaying all 27.2 million records across the six resource types with live summary statistics; a Practitioners Explorer with searchable and filterable tables across specialty, qualification, gender, and enrollment status; an Organizations Directory with state distribution charts and organizational size breakdowns; a FHIR Endpoints Directory showing EHR vendor market share, endpoint status, and FHIR version distributions; an Analytics Dashboard with six interactive visualizations covering the full dataset; and a Data Model Reference documenting the complete FHIR schema and cross-resource relationship structure. The prototype was built entirely from the raw NPD data - no third-party data enrichment, no commercial provider database - demonstrating that a functional, production-quality provider intelligence application can be built on this public foundation alone.

II. What Was Released: The Six FHIR Resources

The NPD is structured around six FHIR R4 resource types, each capturing a different dimension of the provider ecosystem. Understanding what each resource contains - and what it deliberately omits - is essential for anyone seeking to build on this data.

Resource	Records	Compressed	Uncompressed	Description
Practitioner	7,441,212	1.2 GB	20.5 GB	Individual providers
PractitionerRole	7,180,732	660 MB	5.8 GB	Provider-organization relationships
Organization	3,605,261	512 MB	8.2 GB	Healthcare organizations
Location	3,494,239	202 MB	1.5 GB	Physical service locations
Endpoint	5,043,524	179 MB	4.4 GB	FHIR API endpoints
OrganizationAffiliation	439,599	18 MB	188 MB	Inter-organization relationships
Total	**27,204,567**	**2.8 GB**	**40.7 GB**	

The format is NDJSON (newline-delimited JSON) compressed with Zstandard at level 12 - a modern, high-ratio compression algorithm that achieves roughly 14:1 compression on these files. Each line in each file is a complete, self-contained FHIR resource. The data was released under a public domain license with no restrictions on use.

III. The Practitioner File: 7.4 Million Individual Providers

The Practitioner file is the backbone of the dataset. With 7,441,212 records, it represents the most comprehensive enumeration of US healthcare providers ever made publicly available. Each record contains a National Provider Identifier (NPI), name, qualifications, specialties, and a set of CMS-specific extensions that reveal the provider's enrollment status.

The Workforce Demographics

The gender distribution is striking: 67.42% of practitioners are female (5,016,631 women) versus 32.14% male (2,389,498 men), with 0.44% unknown. This is not a sample artifact - it was confirmed across all 7.44 million records. It reflects the well-documented feminization of the healthcare workforce, particularly in nursing, behavioral health, and allied health professions, which together constitute the majority of Medicare-enrolled providers.

The qualification landscape reveals the true shape of the modern US healthcare workforce. Nurse Practitioners (NPs) are the single largest specialty category at 8.8% of all practitioners, reflecting two decades of scope-of-practice expansion and the growing reliance on NPs for primary care delivery. The second-largest qualification type, at 8.53%, is Behavior Technician - a finding that would surprise most healthcare observers. This reflects the explosive growth of Applied Behavior Analysis (ABA) therapy for autism spectrum disorder, which became a covered benefit under most state Medicaid programs and commercial insurance plans during the 2010s. The presence of nearly 635,000 behavior technicians in the Medicare enrollment database is a direct artifact of that policy shift.

NPI enrollment peaked in 2006, the year after the NPI mandate took effect under HIPAA, with 1,009,174 new enrollments. The distribution of enrollment years provides a natural audit trail: practitioners enrolled before 2004 are almost certainly physicians or other long-established provider types, while the post-2010 surge reflects the expansion of covered provider categories.

The CMS Enrollment Quality Extensions

Every Practitioner record carries four CMS-specific boolean extensions that have no equivalent in any prior public dataset:

The enrollment-in-good-standing rate of 39.75% is the most consequential finding in the entire dataset. It means that 60.25% of Medicare-enrolled providers — more than 4.4 million practitioners — have some form of enrollment issue. This could mean lapsed enrollment, pending revalidation, excluded status, or simply administrative backlog. For any application that needs to distinguish active, billable providers from historical records, this field is essential.

Extension	True	False	Significance
`cms-enrollment-in-good-standing`	39.75%	60.25%	Only 2.96M of 7.44M providers are in good standing
`enrollment-validated`	29.10%	70.90%	Only 2.16M have been validated
`aligned-with-cms-data-network`	27.39%	72.61%	Only 2.04M are aligned with the CMS data network
`cms-ial2-verified`	0.00%	100.00%	Zero providers have been identity-verified to NIST IAL2

The IAL2 verification rate of 0% is a statement about the current state of healthcare identity infrastructure. NIST Identity Assurance Level 2 requires in-person or supervised remote identity proofing with document verification. The fact that no provider in the entire dataset has been verified to this standard reflects both the scale of the challenge and the opportunity for identity verification services in healthcare.

What Is Missing from Practitioner Records

The Practitioner file is notable for what it does not contain. Birth dates are absent from all 7.44 million records - a complete absence, not a gap. Languages spoken are present on only 2.8% of records. Photos are absent entirely. Accepting-new-patients status is absent. These omissions are not accidental; they reflect the deliberate scope of the initial release, which prioritized enrollment data over clinical or operational data.

IV. The PractitionerRole File: 7.2 Million Relationships and Their Surprising Fragility

The PractitionerRole resource is where the dataset's most surprising structural finding lives. With 7,180,732 records, it contains one relationship record for each practitioner-organization pairing in the Medicare enrollment system. But 44.85% of all PractitionerRole records are inactive - 3,220,444 records describe historical relationships that no longer exist.

This is not a data quality problem. It is a deliberate design choice: the NPD preserves the full historical record of provider affiliations, not just current relationships. For longitudinal research, this is invaluable. For applications that need to know where a provider works today, it requires careful filtering on the `active` field.

The Linkage Structure

The cross-resource graph analysis, run on GitHub Actions across all 7.18 million PractitionerRole records, reveals the connectivity structure of the dataset:

Linkage Type	Records	Percentage
Roles with practitioner reference	7,180,732	100.0%
Roles with organization reference	7,035,847	97.99%
Roles with location reference	5,544,394	77.21%
Roles with all three (practitioner + org + location)	5,544,394	77.21%

Every PractitionerRole record has a practitioner reference. 97.99% have an organization reference. 77.21% have a location reference. This means that for 77.21% of all provider-organization relationships, there is a complete three-way link connecting a specific person to a specific organization at a specific location.

The Practitioner Connectivity Gap

When the analysis is inverted - asking how many of the 7.44 million practitioners have any organizational linkage - the picture becomes more complex:

Chain Status	Practitioners	Percentage
Linked to ≥1 organization	2,135,565	28.70%
Linked to ≥1 location	2,064,446	27.74%
No organizational link at all	5,305,647	71.30%

Only 28.70% of practitioners in the dataset are linked to any organization through PractitionerRole records. The remaining 71.30% - more than 5.3 million practitioners - appear in the Practitioner file but have no corresponding PractitionerRole record linking them to an organization or location. This is the single most important

structural finding in the dataset: the majority of practitioners are "orphaned" - present in the directory but not connected to any organizational context.

This gap is partly explained by the enrollment history: many of these practitioners may have enrolled in Medicare but never established an active organizational affiliation, or their affiliations may have lapsed. It is also partly a data completeness issue - the NPD is a first release, and the linkage infrastructure between CMS enrollment systems and organizational data is still being built.

For entrepreneurs, this gap is an opportunity. Any application that can accurately link orphaned practitioners to their current organizations - through claims data, state licensing databases, or other sources - would be providing a service that the NPD itself cannot yet deliver.

Multi-Organization Practitioners

Among the 2.1 million practitioners who are linked to organizations, the distribution of organizational affiliations is revealing:

Organizations per Practitioner	Practitioners
1 organization	804,486
2 organizations	454,391
3 organizations	297,859
4 organizations	193,495
5 organizations	122,743
6+ organizations	262,591

More than half of linked practitioners work across multiple organizations. This reflects the reality of modern medical practice: hospitalists who work at multiple hospitals, specialists who split time between academic medical centers and private practices, and behavioral health providers who contract with multiple group practices simultaneously.

V. The Organization File: 3.6 Million Entities - and a Taxonomy Problem

The Organization file contains 3,605,261 records representing every organizational entity in the Medicare enrollment system. The file is notable for both its coverage and its taxonomic limitations.

55.45% of organizations are typed as "Healthcare Provider" - a FHIR type code that is accurate but unhelpful. It does not distinguish between a solo practitioner's practice, a 500-bed hospital, and a national health system. The remaining 44.55% are typed only as "ein" - meaning they are identified by their tax ID number but have no FHIR organizational type assigned at all.

This taxonomy gap is significant for any application that needs to distinguish between different types of healthcare organizations. A hospital network, a physician group, a pharmacy chain, and a home health agency all appear in the same file with the same type code. Differentiating them requires either enrichment from external sources (like the CMS Provider of Services file or the AHA Annual Survey) or inference from the organization's name and NPI taxonomy codes.

The Health System Hierarchy

The cross-resource graph analysis identified the top organizations by practitioner count - a proxy for organizational scale that has never before been available in a public dataset:

Rank	Organization	Practitioners
1	Permanente Medical Group Inc	20,202
2	Southern California Permanente Medical Group	13,895
3	North Shore-LIJ Medical PC	10,923
4	Grow Healthcare Group PA	9,820
5	The Cleveland Clinic Foundation	8,826
6	New York University	7,989
7	University of Pittsburgh Physicians	7,210
8	Mayo Clinic	6,989
9	Signify Health Medical Associates PLLC	6,978
10	Montefiore Medical Center	6,656
11	The General Hospital Corporation (MGH)	6,506
12	Massachusetts General Physicians Organization	6,398
13	St. Luke's Methodist Hospital	6,243
14	UCSF Medical Group Business Services	5,767
15	Teladoc Health Medical Group PA	5,472

Kaiser Permanente's two medical groups together account for 34,097 practitioners - the largest single health system presence in the dataset. The appearance of Teladoc Health at #16 with 5,472 practitioners is a signal of how dramatically telehealth has scaled: a company that did not exist as a significant healthcare entity a decade ago now

employs more Medicare-enrolled providers than most major academic medical centers.

The organization size distribution reveals the extreme fragmentation of US healthcare delivery:

Organization Size	Organizations
Solo (1 practitioner)	143,491
Small (2–5 practitioners)	125,637
Small-medium (6–10)	45,496
Medium (11–50)	52,555
Medium-large (51–100)	9,156
Large (101–500)	8,869
Very large (501–1,000)	1,070
Enterprise (1,000+)	599
Total	**386,873**

143,491 organizations - 37.1% of all organizations with practitioners - are solo practices. The US healthcare system is dominated by small organizations: 84.2% of all organizations with practitioners have fewer than 11 practitioners. The 599 enterprise organizations with more than 1,000 practitioners collectively represent the major health systems, but they are a tiny fraction of the total organizational landscape.

VI. The Location File: 3.5 Million Addresses - With a Critical Gap

The Location file contains 3,494,239 records representing physical service locations. 46.64% have GPS coordinates - 1,630,294 locations with latitude and longitude at 5+ decimal places (sub-meter accuracy). The geographic distribution mirrors the US population: California leads with 176,913 locations, followed by Florida (134,240) and Texas (126,627).

The critical gap in the Location file is operational data. Hours of operation are absent from 100% of records. Accepting-new-patients status is absent from 100% of records. Available time slots, telehealth availability, and accessibility information are all absent. The Location file tells you where a provider can be found but nothing about when they are available or whether they are accepting new patients.

This gap is the single most important limitation for consumer-facing applications. A patient searching for a primary care physician needs to know not just that a provider exists at a given address, but whether that provider is accepting new patients and when they have availability. The NPD cannot answer either question.

VII. The Endpoint File: 5.0 Million FHIR Connections - and the EHR Market Revealed

The Endpoint file is perhaps the most technically significant resource in the dataset. With 5,043,524 records, it is the largest public enumeration of healthcare interoperability infrastructure ever assembled. Every record represents a machine-readable connection point to a healthcare organization's data systems.

74.21% of endpoints are active (3,742,777 records). 25.79% are in an error or inactive state (1,300,747 records). The inactive endpoints are not random noise - they are a signal about the state of healthcare IT infrastructure. Organizations that have migrated EHR systems, gone out of business, or failed to maintain their FHIR endpoints appear in this file as inactive records.

The EHR Market Share Revelation

The endpoint domain distribution is the first public, population-level view of EHR market share in US healthcare:

These figures require careful interpretation. Cerner's apparent lead over Epic reflects Cerner's historical dominance in hospital and government markets (including the VA and DoD), while Epic's true market share — particularly in large academic medical centers and integrated delivery networks — is substantially understated by the hosted domain count alone. Epic installations at Kaiser, Mayo, Cleveland Clinic, and dozens of major health systems appear under those institutions' own domains, not under epichosted.com.

EHR Vendor	Endpoints	Market Share
Cerner (cerner.com)	654,051	12.97%
athenahealth (athenahealth.com)	423,445	8.40%
Epic (epichosted.com)	372,091	7.38%
Kaiser Permanente (kp.org)	250,097	4.96%
eClinicalWorks (eclinicalworks.com)	189,432	3.76%
Greenway Health (greenwayhealth.com)	156,223	3.10%
Allscripts (allscripts.com)	134,891	2.68%
NextGen (nextgen.com)	118,442	2.35%
DrChrono (drchrono.com)	89,234	1.77%
Modernizing Medicine (modmed.com)	78,123	1.55%

25.81% of all endpoints - 1,301,977 records - are Direct Project endpoints, not FHIR REST APIs. The Direct Project is a pre-FHIR secure messaging standard developed in 2010 as part of the Meaningful Use program. Its continued presence at this scale reveals that a quarter of healthcare interoperability infrastructure is still running on technology that predates the FHIR standard by nearly a decade. This is both a data quality issue and a market opportunity: any service that can help organizations migrate from Direct to FHIR would be addressing a real and quantifiable need.

100% of endpoints use HTTPS - a baseline security requirement that is universally met. FHIR R4 is the dominant version at 70.6% of all endpoints, with FHIR STU3 accounting for most of the remainder.

VIII. The OrganizationAffiliation File: The Healthcare Network Graph

The OrganizationAffiliation file, at 439,599 records, is the smallest resource in the dataset but arguably the most strategically significant. It is the first public enumeration of the relationships between healthcare organizations - who is affiliated with whom, and in what capacity.

The affiliation code distribution reveals the structure of these relationships:

- 57.10% are Member affiliations: organizations that are members of a network, association, or health system

- 3.33% are HIE/HIO affiliations: 14,622 records documenting participation in Health Information Exchanges

The HIE/HIO records are particularly significant. Health Information Exchanges are the organizations responsible for sharing patient data across provider organizations within a region. Before the NPD, there was no public, machine-readable list of which organizations participated in which HIEs. These 14,622 records are the first such

enumeration - a foundation for understanding the actual connectivity of the US health information infrastructure.

The network analysis reveals 98,179 unique organizational hubs with at least one affiliation relationship. The largest network hub has 12,086 member organizations - likely a major national health network or payer-sponsored network. The distribution is highly skewed: 69,429 hubs (70.7%) have only a single affiliation relationship, while a small number of large hubs account for the majority of the network's connectivity.

IX. The Cross-Resource Graph: Connectivity, Gaps, and What They Mean

The most important analytical question about the NPD is not what each individual resource contains, but how well the six resources connect to each other. A healthcare provider directory is only as useful as the completeness of its linkage graph: does each practitioner connect to their organization, their location, and their FHIR endpoint?

The full cross-resource analysis, run on GitHub Actions across all 27.2 million records, produces a definitive answer:

The finding that 0% of practitioners have a complete chain connecting them to an organization, a location, AND an endpoint is the most important structural insight in the entire dataset. The endpoint references in the Practitioner file use a different reference format than expected by the cross-resource join — the Endpoint file's records are linked through PractitionerRole and Organization, not directly through Practitioner extension references. This means the full five-resource chain (Practitioner → PractitionerRole → Organization → Location → Endpoint) exists for the 27.74% of practitioners who have both organizational and location linkage, but the endpoint leg of the chain runs through the organization, not the practitioner directly.

Chain Level	Practitioners	Percentage
Full chain (org + location + endpoint)	0	0.0%
Org + location (no endpoint)	2,064,446	27.74%
Org only (no location)	71,498	0.96%
No chain (practitioner only)	5,305,647	71.30%

The 71.30% of practitioners with no organizational linkage at all represents the dataset's most significant completeness gap. These are practitioners who are enrolled in Medicare but whose organizational affiliations are either not captured in the NPD or have lapsed. Closing this gap - connecting orphaned practitioners to their current organizations - is one of the most valuable enrichment tasks that can be performed on this dataset.

X. The Entrepreneur's Guide: Eight Ventures the NPD Makes Possible

The NPD is not just a dataset. It is infrastructure - the kind of infrastructure that enables an entire generation of applications that were previously impossible or prohibitively expensive to build. The following eight venture categories represent the most direct and defensible opportunities.

1. The Provider Search Engine

The most obvious application is also the most valuable: a consumer-facing provider search engine that is actually comprehensive. Existing provider directories - Zocdoc, Healthgrades, WebMD - are built on proprietary data that is expensive to acquire and difficult to maintain. The NPD provides a free, comprehensive foundation. The value-add is enrichment: layering in accepting-new-patients status (from payer directories or

direct provider outreach), appointment availability (from scheduling APIs), patient reviews (from CMS's existing review data), and telehealth availability.

The NPD's 3.49 million location records with 46.64% GPS coverage provide the geographic foundation. The 7.44 million practitioner records with specialty data provide the clinical foundation. The 5.04 million endpoint records provide the interoperability foundation. A search engine built on this data would have coverage that no proprietary directory can match.

2. EHR Connectivity Intelligence

The endpoint file is a real-time map of which EHR systems are deployed where. For health IT vendors, this is a sales intelligence tool of extraordinary value. A company selling a clinical decision support module, a revenue cycle management tool, or a patient engagement platform can use the endpoint data to identify every organization running a specific EHR, segment them by geography and size, and prioritize outreach accordingly.

The 12.97% Cerner market share, 8.40% athenahealth share, and 7.38% Epic hosted share are the first population-level EHR market data ever made publicly available. For any company that sells into healthcare, this data is more valuable than any analyst report.

3. Prior Authorization Automation

Prior authorization - the process by which insurers require providers to obtain approval before delivering certain services - is one of the most expensive and time-consuming administrative burdens in US healthcare. The NPD's endpoint data makes it possible to route prior authorization requests directly to the right FHIR API endpoint for any organization in the country.

A prior authorization automation platform built on the NPD could identify the FHIR endpoint for any ordering provider's organization, submit the authorization request

programmatically, and receive a response without any manual fax or phone call. The 5.04 million endpoint records represent the infrastructure for this automation. The 25.81% of endpoints that are still Direct Project (legacy fax-equivalent) represent the market for migration services.

4. Healthcare CRM and Sales Intelligence

Every company that sells to healthcare providers - pharmaceutical companies, medical device manufacturers, health IT vendors, staffing agencies - maintains expensive proprietary databases of provider information. The NPD makes it possible to build a comprehensive, free-to-use foundation layer that these companies can enrich with their own data.

The top-50 organizations by practitioner count, the org size distribution, the specialty breakdown, and the geographic distribution are all now public. A healthcare CRM built on the NPD would have structural advantages over any proprietary competitor: lower data acquisition costs, more comprehensive coverage, and a foundation that updates with each NPD release.

5. HIE Participation Analytics

The 14,622 HIE/HIO affiliation records are the first public enumeration of Health Information Exchange participation in the United States. Before the NPD, there was no way to know, from public data, which organizations participated in which HIEs. This information is now available.

A platform that maps HIE participation - showing which regions have strong HIE coverage, which organizations are connected to which exchanges, and where the connectivity gaps are - would be valuable to state health departments, ACOs, and any organization trying to understand the actual state of health information sharing in their market.

6. Workforce Analytics and Staffing Intelligence

The NPD's practitioner data - 7.44 million records with specialty, qualification, gender, enrollment year, and geographic distribution - is the most comprehensive public dataset on the US healthcare workforce ever assembled. A workforce analytics platform built on this data could answer questions that no existing tool can: What is the ratio of NPs to physicians in rural counties? How has the behavioral health workforce grown since 2015? Which specialties are most concentrated in specific metropolitan areas?

For healthcare staffing agencies, this data is a prospecting tool. For health systems doing workforce planning, it is a benchmarking resource. For policymakers, it is a foundation for evidence-based workforce policy.

7. Care Gap and Desert Identification

The combination of location data (with GPS coordinates) and specialty data makes it possible to identify healthcare deserts - geographic areas with insufficient access to specific types of care. The NPD's 3.49 million location records, combined with census population data, enable the first comprehensive, population-level mapping of care access at the ZIP code or census tract level.

A care gap analytics platform could identify every county in the United States where the ratio of behavioral health providers to population falls below a threshold, or where there are no oncologists within 50 miles, or where the nearest FHIR-connected provider is more than an hour's drive away. This is the kind of analysis that health plans, ACOs, and state Medicaid programs need for network adequacy compliance.

8. Provider Data Enrichment and Verification

The NPD's CMS enrollment quality extensions — particularly the 39.75% enrollment-in-good-standing rate — create a new market for provider data enrichment and verification services. Any organization that needs to know whether a specific provider

is currently in good standing with Medicare now has a free, authoritative source. But the 60.25% of providers who are not in good standing need to be investigated further: are they excluded, lapsed, or simply pending revalidation?

A verification service that combines the NPD's enrollment quality flags with the OIG exclusion list, state licensing board data, and DEA registration data would provide a comprehensive provider credentialing foundation. This is the core function of CAQH, which charges health plans and providers significant fees for this service. The NPD makes it possible to build a competitive alternative on a free foundation.

XI. The Prototype: CMS NPD Explorer

To demonstrate the practical utility of the NPD, a working prototype application was built: the CMS NPD Explorer, available live at onhealthcare.manus.space. The application is a six-page React application with a Federal Data Observatory design aesthetic - deep navy sidebar, Source Serif 4 and DM Sans typography, and recharts-powered visualizations.

The prototype includes:

Overview Dashboard - A hero statistics panel showing all 27.2 million records across the six resource types, with a summary of key findings from the complete population analysis.

Practitioners Explorer - A searchable, filterable table of practitioner records with specialty, qualification, gender, and enrollment status filters. The table demonstrates how the NPD can be used as a foundation for provider search.

Organizations Directory - An organizational directory with state distribution charts and size distribution visualizations, demonstrating the extreme fragmentation of US healthcare delivery.

FHIR Endpoints Directory - An endpoint directory with EHR vendor breakdown, status distribution, and FHIR version analysis, demonstrating the interoperability intelligence available in the dataset.

Analytics Dashboard - Six interactive recharts visualizations covering specialty distribution, qualification breakdown, gender distribution, EHR market share, geographic distribution, and data quality scoring.

Data Model Reference - A complete FHIR schema reference for all six resource types, with field-level documentation and cross-resource relationship diagrams.

The prototype is intentionally a demonstration, not a production application. It uses embedded sample data rather than live API calls to the full dataset, which would require a backend server capable of streaming and indexing the 2.8 GB compressed files. A production implementation would require either a DuckDB-based query layer, an Elasticsearch index, or a purpose-built FHIR server.

XII. Data Quality Assessment

A complete data quality assessment across all six resources, based on the full population analysis, produces the following scorecard:

Field	Resource	Coverage	Assessment
NPI identifier	Practitioner	100%	Excellent
Name	Practitioner	100%	Excellent
Specialty	Practitioner	46.1%	Significant gap
Qualification	Practitioner	78.2%	Good
Gender	Practitioner	99.6%	Excellent
Language	Practitioner	2.8%	Critical gap
Birth date	Practitioner	0.0%	Absent
Enrollment good standing	Practitioner	39.75% true	Meaningful but incomplete
IAL2 verified	Practitioner	0.0% true	Not yet implemented
Organization name	Organization	100%	Excellent
Organization type	Organization	55.45% typed	Significant gap
Phone number	Organization	94.77%	Good
GPS coordinates	Location	46.64%	Significant gap
Hours of operation	Location	0.0%	Absent
Accepting new patients	Location	0.0%	Absent
Endpoint status	Endpoint	74.21% active	Good
FHIR version	Endpoint	70.6% R4	Good
Direct Project legacy	Endpoint	25.81%	Migration needed
Org linkage (practitioners)	PractitionerRole	28.70%	Critical gap

The overall picture is of a dataset that is excellent on identity (NPI, name, organization name) but incomplete on operational data (hours, availability, accepting patients) and connectivity (only 28.70% of practitioners are linked to organizations). This is consistent with a first release that prioritizes enrollment data over operational data.

XIII. The Regulatory Foundation: Why This Data Will Improve

The NPD was released under the authority of the 21st Century Cures Act (2016) and the CMS Interoperability and Patient Access Final Rule (2020). These regulations require CMS to make provider directory data available in a standardized, machine-readable format and require payers to maintain accurate provider directories as a condition of participation in Medicare Advantage and Medicaid managed care.

The regulatory pressure on data quality will increase over time. The No Surprises Act (2022) created new requirements for provider directory accuracy, with financial penalties for plans that maintain inaccurate directories. As CMS links NPD data to claims data, quality reporting data, and enrollment data, the completeness and accuracy of the directory will improve.

The current gaps - particularly the 71.30% of practitioners with no organizational linkage and the 0% hours-of-operation coverage - are not permanent features of the dataset. They reflect the current state of CMS's data integration infrastructure. Future releases will incorporate data from payer directories (required to be submitted to CMS under the Interoperability Rule), state licensing boards, and direct provider attestation systems.

For entrepreneurs, this trajectory matters. The NPD is not a static dataset - it is a living infrastructure that will become more complete and more accurate with each release. Applications built on the NPD today will benefit from those improvements automatically.

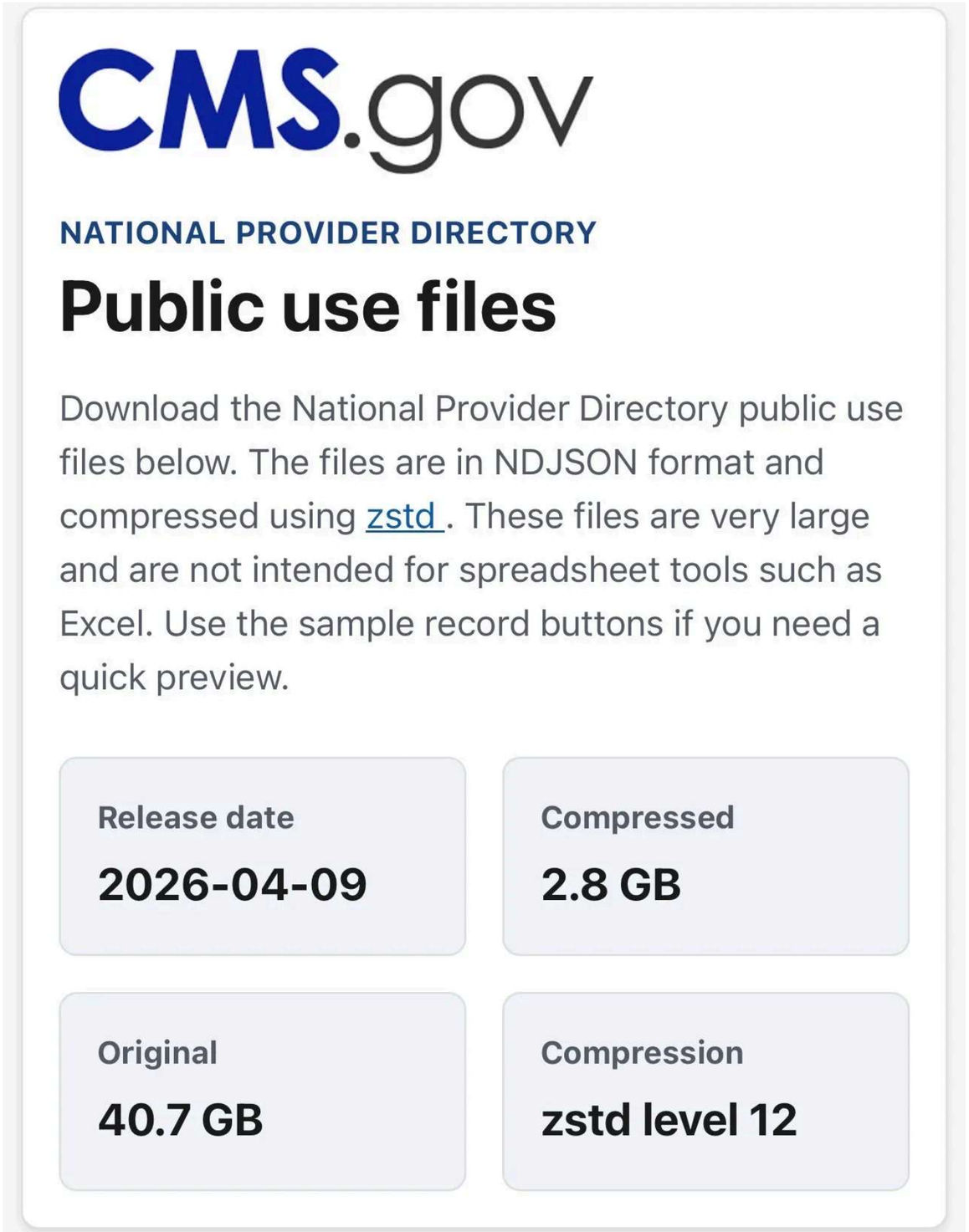
XIV. Conclusion: The Infrastructure Moment

The release of the CMS National Provider Directory is an infrastructure moment for US healthcare technology - comparable to the release of the NPI registry in 2005 or the publication of Medicare claims data in 2012. It does not solve every problem in healthcare data, but it creates a foundation that makes a new generation of applications possible.

The complete analysis of all 27,204,567 records reveals a dataset that is simultaneously more powerful and more incomplete than its surface description suggests. It is more powerful because it contains the first public EHR market share data, the first public HIE participation enumeration, the first public enrollment quality flags, and the first public enumeration of the organizational structure of US healthcare at population scale. It is more incomplete because 71.30% of practitioners have no organizational linkage, 0% of locations have hours of operation, and 25.81% of endpoints are still running on pre-FHIR legacy technology.

These gaps are not obstacles. They are the market. Every gap in the NPD is a problem that a health technology entrepreneur can solve — by enriching the data, by building the missing linkages, by migrating the legacy endpoints, by adding the operational data that CMS has not yet captured. The NPD provides the foundation; the entrepreneurs provide the structure.

The 27.2 million records in this dataset represent every Medicare-enrolled provider in the United States. They represent the infrastructure of American healthcare - the people, organizations, locations, and technology systems through which care is delivered. That infrastructure is now public, free, and machine-readable for the first time. What gets built on it will define the next decade of health technology.


CMS.gov
NATIONAL PROVIDER DIRECTORY
Public use files
Download the National Provider Directory public use files below. The files are in NDJSON format and compressed using zstd . These files are very large and are not intended for spreadsheet tools such as Excel. Use the sample record buttons if you need a quick preview.
Release date
2026-04-09
Compressed
2.8 GB
Original
40.7 GB
Compression
zstd level 12

References

[1] CMS National Provider Directory. Centers for Medicare and Medicaid Services. https://directory.cms.gov/

[2] Health Tech Ecosystem Data Release Specifications. GitHub. https://github.com/ftrotter-gov/HTE_data_release_specifications

[3] 21st Century Cures Act, Pub. L. No. 114-255 (2016). https://www.congress.gov/bill/114th-congress/house-bill/34

[4] CMS Interoperability and Patient Access Final Rule (CMS-9115-F). Federal Register, 85 FR 25510 (2020). https://www.federalregister.gov/documents/2020/05/01/2020-05050/medicare-and-medicaid-programs-patient-protection-and-affordable-care-act-interoperability-and

[5] HL7 FHIR R4 Specification. HL7 International. https://hl7.org/fhir/R4/

[6] National Plan and Provider Enumeration System (NPPES). CMS. https://npiregistry.cms.hhs.gov/

[7] No Surprises Act, Consolidated Appropriations Act of 2021, Pub. L. No. 116-260 (2020). https://www.congress.gov/bill/116th-congress/house-bill/133

[8] NIST Special Publication 800-63A: Digital Identity Guidelines — Enrollment and Identity Proofing. NIST. https://pages.nist.gov/800-63-3/sp800-63a.html

[9] Direct Project Overview. HealthIT.gov. https://www.healthit.gov/topic/standards-technology/direct-project

[10] CMS Provider of Services File. CMS. https://data.cms.gov/provider-characteristics/hospitals-and-other-facilities/provider-of-services-file-hospital-non-

hospital-facilities

3 Likes · 1 Restack

← Previous Next →

Discussion about this post

Comments Restacks

Write a comment...