simple healthcare

Improving Transparency in Coverage Data: Reducing Ghost Rates, Adding Utilization, and Standardizing File Structure

David Muhlestein, PhD, JD

February 2026

Summary

The December 2025 proposed Transparency in Coverage (TiC) updates appropriately target data quality and usability, but the ultimate quality and usability of the data will depend on implementation choices. This paper examines four technical design decisions that will determine whether the proposed rule achieves its stated objectives: ghost rate filtering, adding utilization reporting, accounting for bundled payments and file architecture.

- **Taxonomy-only ghost rate filtering creates systematic errors**. Using Current Procedural Terminology (CPT) code 27447 (knee replacement) as an example, only 32.1 percent of orthopedic surgeons bill for the procedure, while physician assistants account for 33.3 percent of claims. A rule that includes “all orthopedic surgeons” is over-inclusive; a rule that excludes PAs and NPs is under-inclusive. Both errors occur simultaneously under taxonomy-only approaches.

- **Multi-specialty codes resist single-specialty filtering**. For CPT 83540 (iron serum test), no single specialty accounts for more than 30 percent of claims, and the top specialty (hematology-oncology) has only 11.4 percent of its providers performing the code. Any single-specialty inclusion rule will be unstable.

- **Utilization data is necessary but must be structured carefully**. Naive National Provider Identifier (NPI)-level volume counts can double-count the same service event because multiple NPIs appear on a single claim in different roles. Group-level claim counts in the in-network rates file paired with NPI-level binary activity flags in the utilization file provide practical utility without double counting, while being straightforward to create.

- **Bundled payment reporting, particularly in Ambulatory Surgery Centers (ASCs), remains too inconsistent for episode-level comparison.** Expected episode components, such as the price of an implant, are frequently missing from disclosed rates, making it impossible to calculate comparable total costs across plans.

- **Relational rectangular tables are more usable than JavaScript Object Notation (JSON) or flattened comma-separated values (CSV).** Empirical testing shows CSV-based relational outputs are also significantly smaller than JSON file sizes (53 percent on average), with better support for standard analytic workflows.

Specific Recommendations:

1. **Replace taxonomy-only ghost rate filtering with a hybrid approach** combining Taxpayer Identification Number (TIN)-level claims-based inclusion with provider specialty backstops.
2. **Require utilization reporting using group/TIN claim counts** as a primary metric in the in-network rates file, with NPI-level binary activity indicators as supplemental context in the utilization file.
3. **Require explicit bundle/component semantics** for ASC disclosures, including passthrough markers and bundle-component linkage fields.
4. **Require data to be shared in relational rectangular machine-readable files**, such as CSV or Parquet files.

1. Introduction

In December 2025, the Department of the Treasury, Department of Labor, and Department of Health and Human Services (the Departments) released a proposed rule to update Transparency in Coverage requirements. The proposed changes are intended to make machine-readable files more usable by reducing size, improving context, and increasing standardization. That objective is directionally correct.

The remaining issue is technical design. The same policy objective can produce very different results depending on how inclusion logic, utilization logic, and file structure are implemented. A rule that sounds reasonable in regulatory text can produce distorted or unusable data if the implementation details are not carefully specified.

This paper focuses on those technical choices. It is not a full legal response to the proposed rule. Instead, it builds on prior recommendations to provide a practical analytic framework for how to avoid predictable data distortions while preserving feasibility for payers and utility for analysts.

Specifically, this paper addresses four implementation questions:

- How should ghost rates be removed without creating systematic inclusion bias?
- How should utilization be reported so that rates are interpretable and not double counted?
- How can bundled payment rates be appropriately conveyed to accurately express all-in pricing?
- What file structure best balances machine-readability, usability, and analytic reproducibility?

For each question, I present empirical evidence from analysis of current TiC data, identify the failure modes of proposed approaches, and recommend specific alternatives.

2. The Ghost Rate Problem

2.1 The Issue

A major challenge with TiC data is that many reported rates relate to providers and groups that do not provide those services. For example, a psychiatrist will never perform a heart transplant, but insurers may report a negotiated rate for that service. These negotiated rates—known as ghost rates—undermine the goals of transparency data.

In prior work, I analyzed TiC data from 61 insurers including three national commercial insurers (CVS/Aetna, Cigna, and UnitedHealthcare) and 58 Blue Cross Blue Shield plans. Across all insurers, 91.8 percent of all negotiated rates were ghost rates (3,153,469,476 out of 3,433,560,471). At the individual provider level, 95.4 percent of provider-to-billing code pairs were ghost rates.

Ghost rates significantly increase the size of the data and make evaluation more difficult. Researchers and analysts do not know whether to include each reported rate in their analysis, resulting in distrust of the TiC data. The proposed rule's effort to address this problem is appropriate.

2.2 The Proposed Solution and Its Risks

The proposed rule addresses ghost rates through taxonomy-based filtering: excluding provider-code combinations where the provider's specialty taxonomy code does not match codes typically associated with that service. While this would remove some of the ghost rates, taxonomy-only filtering is too coarse because specialty labels are imperfect proxies for whether a provider or group actually furnishes a specific code.

2.3 Two Predictable Errors

Taxonomy-only approaches create two systematic errors simultaneously:

Over-inclusion: Broad specialty inclusion rules keep many provider-code combinations that are technically plausible by taxonomy but uncommon in claims. If a rule includes "all orthopedic surgeons" for knee replacement, it includes the 68 percent of orthopedic surgeons who do not perform knee replacements.

Under-inclusion: Excluding adjacent provider categories (for example, physician assistants (PAs) and nurse practitioners (NPs) in procedural contexts) can remove valid provider-code combinations that are observed in real utilization patterns. If a rule excludes non-physician providers from surgical codes, it excludes the physician assistants who account for one-third of knee replacement claims.

In other words, taxonomy-only filtering can remove the wrong rates and keep the wrong rates at the same time.

2.4 Analysis A: CPT 27447 (Knee Replacement)

CPT 27447 (total knee arthroplasty) is a useful anchor case because it is dominated by orthopedic surgery but still has meaningful participation from PA and NP providers in claims-linked activity.

Table 1 presents three metrics for each specialty:

- **Share of Claims:** The percentage of all CPT 27447 claims attributed to providers in that specialty. This shows which specialties are actually performing the procedure in practice.

- **Share of Providers:** The percentage of all providers who billed for CPT 27447 that belong to each specialty. This indicates the distribution of active providers across specialties.

- **% of Specialty Performing Code:** The percentage of all providers within a given specialty who billed for CPT 27447 at least once. This reveals how common it is for members of that specialty to perform the procedure—a low percentage indicates most providers in that specialty do not perform this code.

Table 1: Claim Share and Provider Participation for CPT 27447

Specialty	Share of Claims	Share of Providers	% of Specialty Performing Code
Orthopedic Surgery	60.8%	57.6%	32.1%
Physician Assistant	33.3%	35.9%	4.3%
Nurse Practitioner	4.6%	4.9%	0.3%
Other Specialty	1.3%	1.7%	—

Source: Author's analysis of Medicare Provider Utilization and Payment Data.

This pattern demonstrates both errors simultaneously. **Over-inclusion from orthopedics-only rules**: Only 32.1 percent of orthopedic surgeons (6,610 out of 20,596) bill for CPT 27447. Including all orthopedic surgeons includes 13,986 providers who do not perform the procedure. **Under-inclusion from excluding PA/NP**: Physician assistants account for 33.3 percent of claims, yet only 4.3 percent of all PAs (4,114 out of 95,002) perform the code. Nurse practitioners account for 4.6 percent of claims, with only 0.3 percent of NPs (557 out of 174,743) performing the code. A rule that excludes non-physicians from surgical codes removes providers responsible for more than one-third of actual utilization.

Table 2: Providers Performing CPT 27447 vs. Total Providers by Specialty

Specialty	Providers Performing Procedure	Total Providers in Specialty	Participation Rate
Orthopedic Surgery	6,610	20,596	32.1%
Physician Assistant	4,114	95,002	4.3%
Nurse Practitioner	557	174,743	0.3%

Source: Author's analysis of Medicare Provider Utilization and Payment Data.

The rate-count sensitivity by filtering method is substantial. Different inclusion rules produce materially different rate counts for the same procedure and payer. Table 3 compares how many negotiated rates would be retained under different filtering approaches:

- **Reported Rates:** The total number of negotiated rates for CPT 27447 published in TiC files by each payer (before any filtering).

- **NPIs in Claims (% of reported):** The number of rates associated with NPIs that actually billed for CPT 27447 in Medicare claims data—this represents a claims-based filtering approach. The percentage shows what share of reported rates would remain after filtering.

- **Orthopedic-Only Inclusion (% of reported)**: The number of rates retained if only orthopedic surgeons are included (taxonomy-based, single specialty). The percentage shows what share of reported rates would remain.

- **Orthopedic + PA/NP Inclusion (% of reported):** The number of rates retained if orthopedic surgeons, physician assistants, and nurse practitioners are all included (taxonomy-based, expanded specialties). The percentage shows what share of reported rates would remain.

Table 3: CPT 27447 Rate Counts by Inclusion Method and Payer

Payer	Reported Rates	NPIs in Claims (% of reported)	Orthopedic-Only Inclusion (% of reported)	Orthopedic + PA/NP Inclusion (% of reported)
Aetna	31,368	4,982 (15.9%)	5,638 (18.0%)	10,067 (32.1%)
Cigna	15,399	2,073 (13.5%)	1,895 (12.3%)	5,446 (35.4%)
UnitedHealthcare	46,352	4,101 (8.8%)	5,993 (12.9%)	19,246 (41.5%)

Source: Author's analysis of Transparency in Coverage data.

These are not small differences at the margin. Claims-based filtering retains only 9–16 percent of reported rates, while taxonomy-based approaches retain 12–42 percent depending on how broadly specialties are defined. The choice of inclusion method changes how many rates are considered relevant by factors of two to five.

2.5 Analysis B: CPT 83540 (Iron Serum Test)

CPT 83540 demonstrates a different version of the same problem: broad multi-specialty distribution with low within-specialty participation.

Table 4 uses the same metrics as Table 1, applied to CPT 83540. Share of Claims and % of Specialty Performing Code are defined as before.

Table 4: Claim Share by Specialty for CPT 83540

CPT 83540 demonstrates a different version of the same problem: broad multi-specialty distribution with low within-specialty participation.

Table 4 uses the same metrics as Table 1, applied to CPT 83540. Share of Claims and % of Specialty Performing Code are defined as before.

Specialty	Share of Claims	% of Specialty Performing Code
Hematology-Oncology	29.8%	11.4%
Internal Medicine	19.0%	2.6%
Pathology	17.0%	0.4%
Family Practice	9.7%	2.2%
Medical Oncology	7.6%	9.4%
Nurse Practitioner	5.2%	0.6%
Nephrology	3.2%	3.6%
Physician Assistant	1.9%	0.5%
Endocrinology	1.9%	1.6%
Rheumatology	1.0%	2.3%

Source: Author's analysis of Medicare Provider Utilization and Payment Data.

No single specialty accounts for more than 30 percent of claims. The top specialty (hematology-oncology) has only 11.4 percent of its providers performing the code. This is exactly where single-specialty or narrow taxonomy approaches become unstable: no single-specialty rule captures real code execution patterns well.

Rate counts again shift materially by method. Table 5 uses a similar structure to Table 3:

- **Reported Rates**: Total negotiated rates published for CPT 83540 by each payer.
- **NPIs in Claims (% of reported):** Rates associated with NPIs that actually billed for this code (claims-based approach). The percentage shows what share of reported rates would remain after filtering.
- **Rates If Common Specialties Included (% of reported):** Rates retained if all specialties commonly associated with this code are included (taxonomy-based approach with multiple specialties). The percentage shows what share of reported rates would remain.

Table 5: CPT 83540 Rate Counts by Inclusion Method and Payer

Payer	Reported Rates	NPIs in Claims (% of reported)	Rates If Common Specialties Included (% of reported)
Aetna	23,338	2,563 (11.0%)	22,224 (95.2%)
Cigna	2,722	112 (4.1%)	2,141 (78.7%)
UnitedHealthcare	43,450	2,073 (4.8%)	36,855 (84.8%)

Source: Author's analysis of Transparency in Coverage Data.

If all common specialties were included using taxonomy-based inclusion criteria, most ghost rates would still be published, preserving the same core challenges. Taxonomy-based filtering retains 79–95 percent of reported rates across these payers, while claims-based filtering retains only 4–11 percent. For Aetna, 95 percent of all negotiated rates would be included under taxonomy filtering, even though with the claims-based approach, only 11 percent of reported rates reflect providers with a billing history for that code.

2.6 Recommended Hybrid Method for Ghost Rate Removal

A better approach combines claims evidence with specialty context.

Rule 1: TIN Evidence Window Include provider groups if the group TIN appears on one or more claims for the billing code in a 12-month period ending six months before the quarterly file posting.

Rule 2: Specialty Connection Backstop Permit specialty-based inclusion if any NPI in the group has a specialty strongly connected to the billing code based on claims patterns. For example, this threshold might be set at 20 percent, where if 20 percent or more of each specialty bill for that service, then all providers with that specialty who have a negotiated rate would be included. The specific threshold percentage could be determined by CMS.

Rule 3: No Minimum Threshold for Inclusion One-or-more claim appearances triggers inclusion in the lookback window. This avoids arbitrary volume thresholds that may exclude legitimate low-volume providers.

This method keeps the operational simplicity of a rules-based filter but anchors inclusion to observed billing activity. It reduces both types of classification error relative to taxonomy-only approaches.

The specialty backstop addresses a practical limitation of claims-only filtering: new providers who have not yet billed for a service but are likely to do so. Including specialties where a high proportion of providers perform the service captures these prospective providers while limiting the inclusion of true ghost rates. The specific threshold for specialty inclusion could be established through rulemaking.

2.7 Implementation Feasibility

This approach is feasible because it uses fields and relationships already present in payer workflows: group-level identifiers, provider identifiers, and claims history. Payers already maintain claims data for their own operations; using that data to inform inclusion decisions does not require new data collection.

The approach can also be audited and replicated across payers in a standardized way. CMS could specify the lookback window, the minimum claim threshold (one or more), and the specialty-connection logic, and payers could implement consistent filtering without substantial new infrastructure.

3. Volume and Utilization: Why It Matters and How to Avoid Counting Errors

3.1 Why Utilization Is Necessary for TiC Usefulness

Rate files without utilization context are difficult to interpret. A posted rate may be technically valid and still have minimal practical relevance if there is little observed service activity.

Utilization data enables several core uses:

- **Proxy for relevance:** Higher observed claim activity indicates rates tied to active delivery patterns, which helps prioritize analysis.

- **Proxy for quality**: Higher volumes of care have been shown to be strongly correlated with quality (Gooiker et al., 2011; Post et al., 2010; Rafaqat et al., 2024).

- **Network development:** Plans and purchasers can identify which groups are actually performing specific services.

- **Competitive analysis:** Analysts can distinguish active networks (which groups are actively performing which services) from groups that are only nominally in the network.

- **Procedure concentration analysis**: Stakeholders can see where services are concentrated across groups and markets.

- **Rate negotiations:** Using volumes to create weighted average market benchmarks to inform payer/provider rate negotiations.

The proposed rule appropriately recognizes that volume data has value, but it does not currently require direct volume reporting. This paper recommends going beyond the current proposal: **actual claims-based counts of services should be reported in the in-network file at the provider group level, while a binary participation flag should be included for individuals at the NPI level**.

3.2 Why Claims Structure Can Create Counting Errors

The same claim can contain multiple identifiers and roles, and those roles vary by claim type and submission context. Professional and facility claims can include combinations of billing organization identity, provider NPIs in different roles (billing, performing, referring, supervising), and place-of-service context.

If utilization is counted naively at the NPI level, the same service event can be represented multiple times because:

- Multiple NPIs may appear on a single claim

- Provider roles (billing vs. performing vs. referring) are not interchangeable

- Claim-header and line-level fields may point to different NPIs

That creates inflated or unstable provider-level volume estimates. Table 6 illustrates how a single claim can list multiple NPIs in different roles:

- **Field:** The type of NPI field on the claim (billing, performing, referring, or supervising).

- **Value:** A sample NPI number.

- **Naive Count Attribution**: What happens if volume is attributed to every NPI on the claim—each gets +1, leading to quadruple-counting of a single service event.

Table 6: Example of Multi-Counting Risk from a Single Claim

Field	Value	Naive Count Attribution
Billing NPI	1234567890	+1 to NPI volume
Performing NPI	0987654321	+1 to NPI volume
Referring NPI	1111111111	+1 to NPI volume
Supervising NPI	2222222222	+1 to NPI volume

A single service event could be counted four times if volume is attributed to each NPI on the claim

3.3 Recommended Utilization Reporting Model

Utilization should be reported at two different levels. First, the group-level utilization should be reported in the in-network file. Second, a binary participation flag should be provided at the NPI-level in the utilization file.

In-Network Rates File

Unit definition: number of claims where the group TIN appears as the billing entity. This would be reported at the provider group level and would include actual claims. Reporting bins could follow established CMS cell size suppression policy as used in Medicare Advantage monthly enrollment data:

Table 7 defines the recommended reporting categories for utilization counts:

- **Range**: The claim count category.

- **Description:** How that range should be reported in the file and what it indicates about provider activity.

Table 7: Recommended Utilization Reporting Bins

Range	Description
0	No claims in lookback period, "0" reported
10 or fewer	Low volume (exact count suppressed), "10 or fewer" reported
11+	Exact value reported

Groups with no historical claims should only be included when taxonomy indicates they are likely to perform the service in the future. A 0 / 10 or fewer / 11+ exact structure is consistent with established federal public reporting patterns that suppress very small cells while still enabling analysis. Zero claims pose no privacy risk because privacy protections are focused on patients, not providers. The "10 or fewer" category indicates that a provider group performs the procedure while avoiding false precision that may reflect anomalies. As a recommendation, all volumes should be reported at the network level, not the plan level.

Utilization File

The proposed utilization file should be adopted and should be distinguished from the in-network data by only including binary flags about whether each NPI appeared at least once as performing provider in the lookback window for various codes and places of service. This is intentionally not an NPI-level claim count. This structure preserves practical utility of identifying the specific providers who perform each service, while avoiding the most common overcounting errors.

3.4 Practical Implementation Guide

The intended use is:

- Treat group claim counts as the primary utilization signal
- Treat NPI binary indicators as provider-level activity confirmation
- Avoid inferring exact provider-level volumes from mixed-role claims fields

Table 8 maps common analytic use cases to the appropriate utilization metric:

- **Use Case:** The analytic objective.
- **Why Volume Matters**: The reason utilization data is relevant for this use case.
- **Risk with Naive NPI Counts**: The problem that occurs if NPI-level counts are used without role attribution.
- **Recommended Metric:** The utilization measure that avoids the identified risk.

Table 8: Recommended Utilization Outputs by Use Case

Use Case	Why Volume Matters	Risk with Naive NPI Counts	Recommended Metric
Quality/relevance proxy	Distinguish nominal from actively used rates	Role-level double counting	TIN claim count + NPI binary
Network development	Identify active service providers	Misattributed provider activity	TIN claim count
Competitive analytics	Compare effective market activity	Inflated provider-level counts	TIN claim count + bins

3.5 Feasibility

This approach does not require new data infrastructure. Payers already reference claims data to produce the allowed amount files for out-of-network services. Counting claims at the group/TIN level is straightforward using existing claims processing systems. The proposed move from monthly to quarterly reporting reduces the frequency of file generation; while adding utilization data increases the scope of each file, the net operational burden would likely be comparable to or less than current requirements.

One limitation is that utilization pattern examples in this paper rely primarily on Medicare-based reference data, while the policy applies to commercial populations. However, the structural patterns—multi-specialty distribution, low within-specialty participation rates, and multi-role claim attribution—are characteristics of claims processing that apply across payers. The specific percentages may vary by market, but the directional findings and the recommended approach remain valid for commercial data.

4. Bundles and Passthrough Costs

Bundled payments, particularly those performed at ASCs, remain inconsistent across files and contracts. In knee replacement examples, rates can appear as fully itemized components, partially bundled structures, bundled with passthrough elements, or partial sets where key facility components are absent. This inconsistency prevents reliable episode-level comparison.

4.1 Core Technical Issues

Missing facility components produce incomplete episode totals. An expected knee replacement episode includes facility fees, surgeon professional fees, assistant surgeon fees, anesthesia, implant/device costs, imaging, and medications. When facility or implant components are absent from disclosed rates, users cannot calculate a total episode cost.

Bundled and unbundled rates are mixed without explicit machine-readable linkages. A rate may represent a bundle that includes multiple services, or it may represent a single component that should be added to other rates. Without explicit flags, users cannot determine which interpretation applies.

Passthrough device/medication treatment is inconsistent and often implicit. Some rates include device costs; others treat devices as passthroughs billed separately through an invoice. The distinction is not always machine-readable.

4.2 Core Technical Issues

Table 9 contrasts what a complete knee replacement episode should include versus what is typically available in TiC data. The table presents two scenarios side-by-side:

- **Code:** The CPT, HCPCS, or other billing code for each service component.
- **Modifier**: Any modifier applied to the code (e.g., modifier 80 for assistant surgeon).
- **Description**: A brief description of the service.
- **Category:** The type of service (facility fee, professional fee, anesthesia, implant, etc.).
- **Price**: The negotiated rate for that component.

Table 9: Expected vs. Available ASC Knee Replacement Components

Expected Bundle (Complete Episode):

Code	Modifier	Description	Category	Price
27447	—	Total knee arthroplasty	Facility Fee	$18,500
27447	—	Total knee arthroplasty	Surgeon Professional	$3,200
27447	80	Total knee arthroplasty - assistant surgeon	Assistant Surgeon	$650
01402	—	Anesthesia for total knee replacement	Anesthesia	$1,100
C1776	—	Joint device (implantable)	Implant/Device	$5,500
76942	—	Ultrasonic guidance for needle placement	Imaging	$150
64447	—	Femoral nerve block	Anesthesia Add-on	$425
J2001	—	Lidocaine injection	Medication	$18
J1100	—	Dexamethasone injection	Medication	$32
J2405	—	Ondansetron injection	Medication	$25
			Total	**$29,600**

Example of Available Data (Incomplete):

Code	Modifier	Description	Category	Price
27447	—	Total knee arthroplasty	Surgeon Professional	$3,200
01402	—	Anesthesia for total knee replacement	Anesthesia	$1,100
76942	—	Ultrasonic guidance for needle placement	Imaging	$150
64447	—	Femoral nerve block	Anesthesia Add-on	$425
			Total	**$4,875**

Source: Author's analysis of Transparency in Coverage data. Example is illustrative of patterns observed across multiple payers.

The available data total ($4,875) is 16 percent of the expected episode total ($29,600). The missing components—facility fee and implant—represent the majority of episode cost. A user comparing "knee replacement prices" across plans using only disclosed rates would be comparing incomplete and non-comparable figures.

Similar challenges occur with hospital inpatients. While many hospitals do receive commercial payments based on diagnosis related groups (DRGs) that function similarly to Medicare, research using price transparency data shows that only about 29 percent of hospitals include DRGs in all their contracts, while others are paid by a combination of revenue center (RC) and miscellaneous CPT codes, making it very difficult to assess the likely cost of an admission.

4.3 Why This Matters

The core goal of price transparency is enabling comparison. Without explicit bundle semantics, users cannot determine what is included in a displayed episode price, whether component lines should be added or treated as informational, or whether two plans are disclosing equivalent package definitions. Incomplete or inconsistent bundle disclosure makes cross-plan comparison impossible for exactly the high-cost procedures where comparison matters most.

4.4 Directional Standardization Approach

Require bundle-oriented structure with:

- **Bundle identifier:** A unique ID linking all components of a bundled episode
- **Bundle total:** When applicable, the total price for the bundle
- **Component records with inclusion flags**: Each component code with a flag indicating whether it is included in the bundle total or billed separately
- **Passthrough markers:** Explicit indication of which components are passthroughs (billed at cost, not included in bundle) and what those prices typically are
- **Bundle-component linkage:** Explicit foreign key relationship between component lines and bundle totals
- **Average paid amounts:** Include average paid amounts for various bundles

4.5 Looking Forward

Bundled payments are not paid consistently between insurers, reflecting underlying variation in how episodes are contracted. Standardizing bundle reporting requirements would create pressure for payers and providers to negotiate contracts that align with those requirements. Over time, this could reduce variation in how bundled payments are structured, making both reporting and price comparison more tractable. Transparency rules that specify clear bundle semantics may therefore have effects beyond disclosure—they can influence how future contracts are written.

5. Data Architecture: Relational Rectangular Tables Are Better Than JSON or Flat CSV Files

5.1 Problem Statement

The current proposed rule recognizes limitations with both JSON and CSV (or other rectangular file formats) and discusses pros and cons of each. JSON enables hierarchical relationships but often requires heavy parsing and table reconstruction. A single flattened CSV is easy to open but can collapse key relationships or duplicate data extensively.

However, the framing in the proposed rules is incomplete. CSV files in a relational structure can combine the flexibility of JSON with the ease of use of CSV. The core question is not just serialization format; it is whether the data model is preserved in a way that is machine-readable, reproducible, and usable for typical analytic workflows. A flattened CSV is cumbersome and much too large, while JSON must be transformed into relational tables to be useful.

To illustrate why both current formats create barriers for analysts, the following subsection presents real sample records from UnitedHealthcare files in both JSON and flattened CSV formats. These examples demonstrate the specific structural challenges that a relational approach would resolve.

5.1.1 Illustrative Structure Examples

JSON Sample Excerpt

```
{
 "reporting_entity_name": "United HealthCare Services, Inc.",
 "reporting_entity_type": "Third-Party Administrator",
 "last_updated_on": "2026-02-01",
 "version": "2.0.0",
 "provider_references": [
  {
   "provider_groups": [
    {
     "npi": [1477724250],
     "tin": {
      "type": "ein",
      "value": "205942007",
      "business_name": "PHYSICIANS ROME SURGERY CENTER"
     }
    }
   ],
   "provider_group_id": 25928
  }
 ],
```

JSON Sample Excerpt (cont.)

```
 "in_network": [
 {
    "negotiation_arrangement": "ffs",
    "name": "RBC DNA HEA 35 AG 11 BLD GRP
WHL BLD CMN ALLEL",
    "billing_code_type": "CPT",
    "billing_code": "0001U",
    "description": "Red blood cell antigen
typing, DNA, human erythrocyte antigen gene
analysis of 35 antigens from 11 blood groups,
utilizing whole blood, common RBC alleles
reported",
    "negotiated_rates": [
      {
       "provider_references": [25928],
       "negotiated_prices": [
        {
         "setting": "outpatient",
         "negotiated_rate": 302.4,
         "negotiated_type": "negotiated",
         "billing_class": "professional",
         "expiration_date": "9999-12-31"
        }
       ]
      }
     ]
    }
  ]
}
```

The JSON format stores data in deeply nested structures. Provider information appears in a separate provider_references array at the top of the file, linked to negotiated rates by numeric ID. Specifically, the IDs in each negotiated rate's provider_references array must be matched to provider_group_id records in the top-level provider_references section to identify the associated TINs and NPIs.

To associate a rate with its provider, an analyst must first parse the entire file, build an index of provider references, and then traverse nested arrays to reconstruct each rate record. This parsing overhead is substantial for large files (I have processed files over 200GB), requiring specialized streaming parsers and significant computing resources. Payers construct these JSON files from their own relational databases, and analysts must then reconstruct relational tables to perform any meaningful analysis—an unnecessary round-trip that adds complexity without adding value.

Table 10 shows sample rows from a fully flattened CSV file. Each column represents a data field from the TiC schema; the key observation is that most columns repeat identical values across rows, with only the provider identifiers (TIN, NPI) changing.

Table 10: Flattened CSV Sample Rows

reporting_entity_name	reporting_entity_type	last_updated_on	version	billing_code	billing_code_type	name	negotiated_rate	negotiated_type	billing_class	service_code	tin	npi
United Health Care Services, Inc.	Third-Party Administrator	2026-01-08	V2.0.0	0FY00Z2	ICD	Transplantation of Liver…	209,216	negotiated	institutional	21	746002164	1821011248
United Health Care Services, Inc.	Third-Party Administrator	2026-01-08	V2.0.0	0FY00Z2	ICD	Transplantation of Liver…	209,216	negotiated	institutional	21	650844880	1083644033
United Health Care Services, Inc.	Third-Party Administrator	2026-01-08	V2.0.0	0FY00Z2	ICD	Transplantation of Liver…	209,216	negotiated	institutional	21	952226406	1689608150

The flattened CSV approach solves the parsing problem but creates a different one: massive redundancy. In the example above, each row repeats identical reporting entity, billing code, rate, and billing class information—only the provider (TIN/NPI) differs. When a single rate applies to thousands of providers, the same data is duplicated thousands of times. Fully flattened CSV files for a single network can reach multiple terabytes, making them impractical to store, transfer, or analyze with standard tools.

5.2 Why Relational Rectangular Outputs Are a Better Standard

Relational rectangular publication provides a middle path:

- **Preserves relationships explicitly** through foreign key linkages
- **Supports direct loading** to Structured Query Language (SQL) databases and tabular analysis pipelines
- **Reduces ambiguity in joins** by making relationship cardinality explicit
- **Improves reproducibility** for analysts and regulators
- **Can reduce redundancy** relative to flattened outputs that repeat shared attributes
- **Well understood** by analysts, researchers and tech companies
- **Widely supported** by many free and professional tools

5.3 File Size and Performance Evidence

In addition to providing easier access to the data and facilitating analysis, a relational CSV structure would likely be significantly smaller than the current standard JSON format. Table 11 is a representative sample of Q4 2025 TiC files and reports compressed file sizes only. The conversion pipeline uses the relational schema in Section 5.5.

- **Insurer:** The name of the reporting entity.
- **JSON Size (compressed):** The file size of the original JSON file after gzip compression.
- **CSV Size (compressed):** The combined file size of all relational CSV tables after gzip compression.
- **CSV to JSON:** The CSV size as a percentage of the JSON size—lower percentages indicate greater space savings from the relational format.

Table 11: File Size Comparison of JSON and CSV Files

Insurer	JSON Size (compressed)	CSV Size (compressed)	CSV to JSON
Aetna Life Insurance Company	1,968 MB	991 MB	50%
American Specialty Health	81 MB	22 MB	28%
Anthem Blue Cross and Blue Shield Virginia	362 MB	57 MB	16%
Baylor Scott and White	48 MB	7.4 MB	15%
Blue Cross and Blue Shield of Kansas	2,521 MB	57 MB	2%
Blue Cross and Blue Shield of Minnesota	1.2 MB	0.81 MB	67%

Insurer	JSON Size (compressed)	CSV Size (compressed)	CSV to JSON
Blue Cross and Blue Shield of Nebraska	4.4 MB	2.1 MB	49%
Cigna Health Life Insurance Company	1,608 MB	834 MB	52%
Medical Mutual Of Ohio	2.5 MB	2.2 MB	89%
Molina Healthcare of Mississippi	0.14 MB	0.07 MB	48%
Sagamore Health Network	1,591 MB	48 MB	3%
United HealthCare Services Inc	5.4 MB	3.6 MB	68%
VSP Vision Care Inc	0.08 MB	0.04 MB	58%

Source: Author's analysis of Transparency in Coverage files converted to relational CSV format.

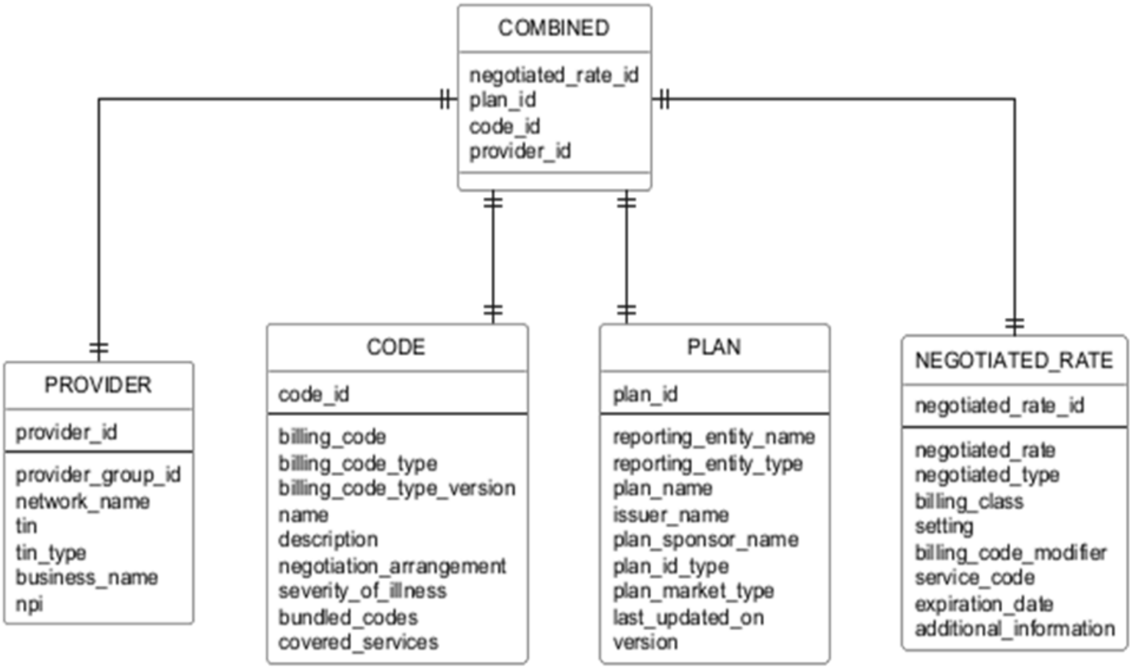
Across 2,249 TiC files, the median compressed CSV files were 44 percent the size of comparable zipped JSON files, and only 12 compressed CSV files were larger than the JSON files. Collectively, compressed CSV files were 53 percent the size of compressed JSON files. The relational CSV format shows significant space savings while preserving all data relationships and is significantly easier to work with.

5.4 Simplified Five-Table Model

I present two potential schemas that could be utilized to share the CSV data in a relational framework.

The simplified standard centers on a combined table and four linked tables: negotiated_rate, plan, code, and provider.

Figure 1: Five-Table Relational Model



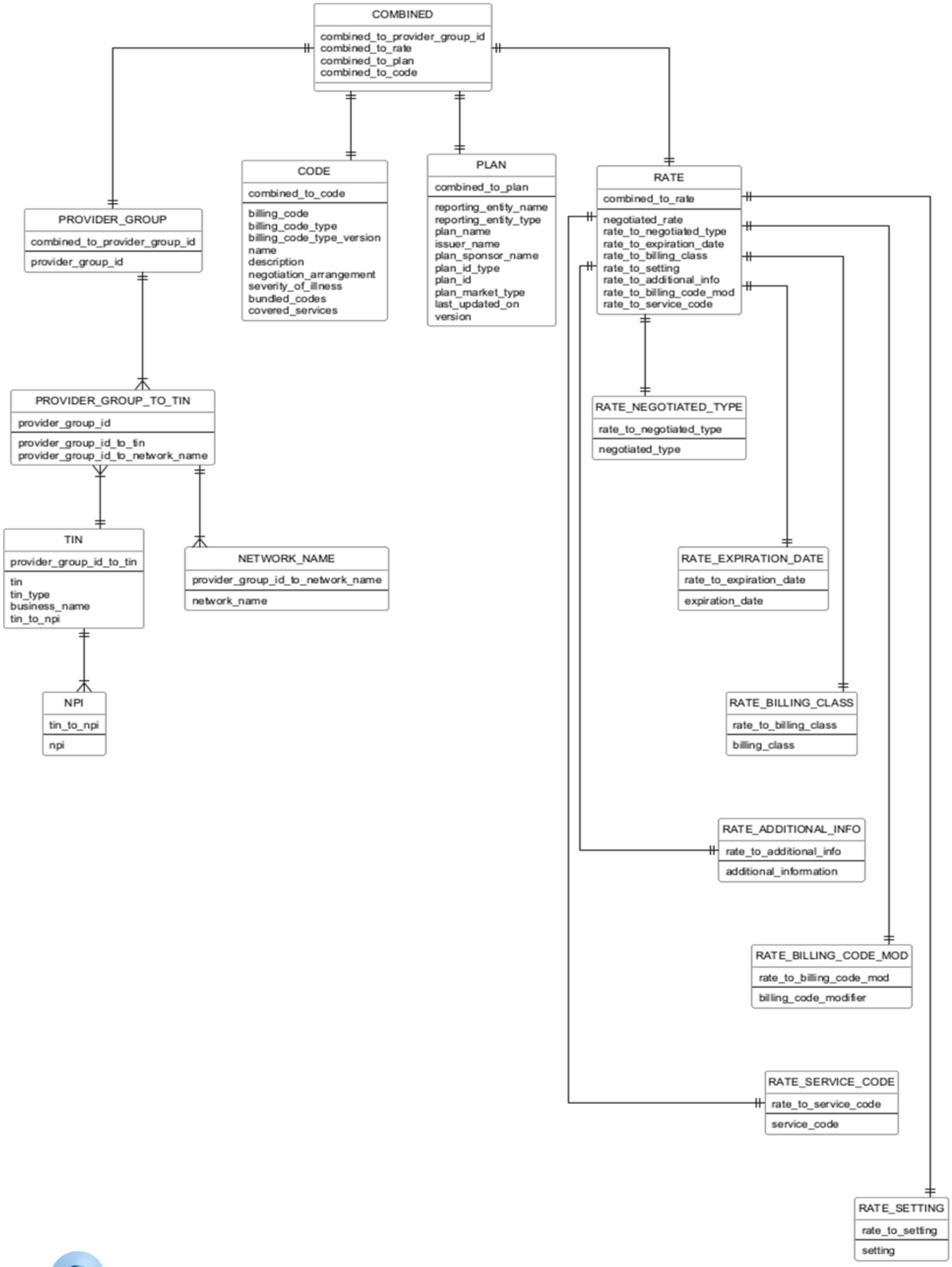
This structure allows users to:

- Query rates by plan, code, or provider without reconstructing nested JSON
- Join tables using standard SQL operations
- Load data directly into analytic tools without custom parsing
- Validate referential integrity through foreign key constraints

5.5 Optional Fully Normalized Variant

For large-scale implementations, selected negotiated rate attributes can be normalized into smaller lookup tables while retaining the same combined linkage strategy. For maximum compression and scalability, the data model can be fully normalized with separate lookup tables for each low-cardinality attribute. This approach is particularly effective for large-scale implementations where storage and transfer costs are significant. This model includes the relationship for data elements in the V2.0 schema. My work and research leverages this fully normalized schema and I have found it to be very efficient for conducting complex analyses.

Figure 2: Fully Normalized Relational Model



This fully normalized structure provides several advantages over the simplified five-table model:

- **Provider hierarchy is explicit:** The provider_group → tin → npi relationship reflects how provider groups are actually constructed in TiC data, where a group contains one or more TINs, each with one or more NPIs.

- **Rate attribute lookup tables reduce redundancy**: Attributes like billing_class, negotiated_type, and expiration_date have limited cardinality (often fewer than 100 distinct values across millions of rates). Storing these as lookup tables with integer foreign keys substantially reduces file sizes.

- **Scalability:** For payers with hundreds of millions of rate records, the storage savings from normalized lookup tables can be substantial.

The tradeoff is query complexity: reconstructing a complete rate record requires joining multiple tables. For most analytic use cases, the simplified five-table model in Section 5.4 provides sufficient structure with lower join overhead.

5.6 Federal Precedent for Relational Public Data

This approach is consistent with existing federal data publication practice. CMS cost report extracts (Healthcare Cost Report Information System, or HCRIS) are distributed as linked multi-file data structures (for example, Rpt, Nmrc, and Alphnmrc tables) and are routinely used in relational workflows. The approach proposed here follows the same design pattern.

5.7 Alternatives to CSV while Maintaining Relational Rectangular Data

While the proposed rules discuss the tradeoff between JSON and CSV files, the Departments do not need to flatly define the file type. Instead, the Departments should require the data to be shared in a standardized relational structure using rectangular data.

When crafting specific implementation and technical rules, different rectangular file types should be considered. For example, Apache Parquet files are becoming much more common and have the advantages of including internal compression, being much faster to query, and interoperable with all modern databases and analytics tools.

For example, I evaluated a single file from Anthem Blue Cross and Blue Shield Colorado. The raw JSON was 10.1GB, and the zipped JSON was 2.1GB. The unzipped relational CSV files were 1.8GB and the zipped CSV files were 124MB. The unzipped (but internally compressed) parquet files were 316MB and the zipped parquet files were 125MB.

5.8 Implementation Feasibility

Transitioning from JSON to relational rectangular formats requires changes to file generation pipelines, but the one-year implementation period following rule adoption provides sufficient time to accomplish this. Payers already structure their internal data relationally; the change is in serialization format, not data model. The relational schemas presented here map directly to the fields already required under existing TiC rules. Conversion tools and libraries for CSV and Parquet are mature and widely available. The primary implementation effort is pipeline engineering, not new data collection or system architecture.

6. Consolidated Recommendations

Table 12 consolidates the recommendations from each section of this paper, organized by category.

Table 12: Summary of Recommendations

Category	#	Recommendation
Ghost Rate Filtering	1	Replace taxonomy-only filtering with a hybrid approach combining TIN-level claims evidence and specialty-connection backstops
	2	Specify a standard lookback window: 12 months of claims data ending six months before the quarterly file posting date
	3	Use a one-or-more threshold: Any claim appearance in the lookback window triggers inclusion; avoid arbitrary minimum volume thresholds
	4	Require method transparency: Payers should publish their inclusion logic and data quality checks so analysts can understand and replicate filtering decisions
Utilization Reporting	5	Use TIN/group claim counts as the primary utilization metric in the in-network rates file
	6	Report utilization at the network level, not the plan level, consistent with the proposed move to network-based reporting
	7	Standardize reporting bins as 0, 10 or fewer, and 11+ with exact values for counts above 10, following established CMS cell size suppression policy
	8	Adopt the proposed utilization file with NPI-level binary flags indicating whether each provider performed each service, rather than NPI-level claim counts

Category	**#**	**Recommendation**
	9	Specify role attribution: Utilization should be attributed based on performing provider role, not billing or referring roles
Bundle Reporting	10	Require explicit bundle/component semantics in machine-readable outputs
	11	Require passthrough indicators distinguishing included components from separately billed items, with typical passthrough prices
	12	Require bundle-component linkage fields connecting individual service lines to bundle totals
	13	Include average paid amounts for bundled episodes to provide context on typical total costs
Data Architecture	14	Require relational rectangular machine-readable standards rather than JSON-only or flattened CSV publication
	15	Publish versioned schemas with stable field names and data types
	16	Consider alternative file formats such as Apache Parquet, which offer internal compression and faster query performance while maintaining relational structure
	17	Provide mapping guidance for converting between formats

The existing price transparency rules have already begun to change how healthcare prices are understood, with researchers, employers, and policymakers using the data to identify price variation and inform decisions. The proposed rule has the potential to significantly expand what is possible.

For employers selecting networks and designing benefit plans, accurate data on which providers actually perform services—and at what price—enables steering toward high-value care. For patients and entrepreneurs creating patient-facing tools, usable price data supports informed decisions about where to seek care. For researchers and regulators, reliable price and volume data enables studies of price variation, market dynamics, and the effects of policy changes. Accomplishing this is much more difficult if the data are filled with ghost rates, lack utilization context, or require custom parsing to access.

The recommendations in this article address specific technical problems that currently limit the data's utility. Claims-based filtering will produce more accurate provider lists than taxonomy-only approaches. TIN-level utilization reporting will give context about the overall activity of the group that NPI-level counts cannot. Explicit bundle semantics will make episode prices comparable across plans. Relational rectangular file formats will make the data accessible to standard analytic tools. These are practical changes that can be implemented within the existing regulatory framework.

Getting the technical details right will determine whether the proposed rule achieves its policy goals.

Author's Note

Support for this work was provided by the Peterson Center on Healthcare. The author is employed by Simple Healthcare, which conducts work with price transparency data.

References

1. Centers for Medicare & Medicaid Services. Transparency in Coverage Proposed Rule (CMS-9882-P). Federal Register. December 23, 2025. https://www.federalregister.gov/documents/2025/12/23/2025-23693/transparency-in-coverage

2. Muhlestein D. High Prevalence of Ghost Rates in Transparency in Coverage Data. Health Affairs Scholar. 2025. https://academic.oup.com/healthaffairsscholar/article/3/11/qxaf212/8321476

3. Muhlestein D. Improving price transparency data: recommendations from practice. Health Affairs Forefront. March 19, 2025. https://www.healthaffairs.org/content/forefront/improving-price-transparency-data-recommendations-practice

4. AAPC. CPT Code 27447 - Total knee arthroplasty. https://www.aapc.com/codes/cpt-codes/27447

5. AAPC. CPT Code 83540 - Iron serum test. https://www.aapc.com/codes/cpt-codes/83540

6. Centers for Medicare & Medicaid Services. Medicare Provider Utilization and Payment Data: Physician and Other Practitioners. https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-provider-and-service

7. Gooiker GA, van Gijn W, Wouters MW, et al. Systematic review and meta-analysis of the volume-outcome relationship in pancreatic surgery. Br J Surg. 2011;98(4):485-494. https://doi.org/10.1002/bjs.7413

8. Post PN, Kuijpers M, Ebels T, Zijlstra F. The relation between volume and outcome of coronary interventions: a systematic review and meta-analysis. Eur Heart J. 2010;31(16):1985-1992. https://doi.org/10.1093/eurheartj/ehq151

9. Rafaqat W, Deng Y, Chen AY, et al. Association of surgeon volume with outcomes for patients with rectal cancer undergoing sphincter-preserving surgery. J Am Coll Surg. 2024;238(5):913-922. https://doi.org/10.1097/xcs.0000000000000913

References

10. Research Data Assistance Center (ResDAC). CMS Cell Size Suppression Policy. https://resdac.org/articles/cms-cell-size-suppression-policy

11. Centers for Medicare & Medicaid Services. Medicare Advantage Monthly Enrollment by Contract. https://www.cms.gov/data-research/statistics-trends-and-reports/medicare-advantagepart-d-contract-and-enrollment-data/monthly-enrollment-contract

12. Chernew ME, Hicks AL, Shah SA. Wide variation in commercial health care prices: implications for price transparency initiatives. Health Aff (Millwood). 2025. https://pubmed.ncbi.nlm.nih.gov/41109326/

13. Centers for Medicare & Medicaid Services. Hospital Cost Report Public Use Files (HCRIS). https://www.cms.gov/data-research/statistics-trends-and-reports/cost-reports/cost-reports-fiscal-year

14. Apache Software Foundation. Apache Parquet. https://parquet.apache.org/

15. CMS Price Transparency Schema. GitHub. https://github.com/CMSgov/price-transparency-guide/tree/develop/schemas/in-network-rates