BCBS Texas

Negotiated Price Transparency

Guide to the Data, How It Was Built, and How to Analyze It

Texas 2036 — Data & Research Team

June 2026; auto updates with each monthly run.

Project folder: `_datashare/Insurance_PriceTransparency/BCBSTX/`

Contents

1 What this project is (and demonstrates for feasibility of this work)

A federal rule called **Transparency in Coverage** requires health insurers to publish, every month, the prices they have negotiated with health-care providers. Blue Cross and Blue Shield of Texas publishes these as very large machine-readable computer files. This project downloads those files, keeps only the parts useful for Texas policy analysis, and turns them into clean datasets of negotiated prices.

The framing follows along in a parallel manner to some of the Texas 2036's existing **hospital** price-transparency work[1]. The core message is the same: identical, shoppable services carry very different negotiated prices, and that variation is information employers and policymakers can use.

The one thing to understand first

Each row of the main dataset (`bcbstx_prices.dta`) is **one distinct negotiated price for one procedure in one provider network**. A price that the insurer gives to thousands of providers is collapsed into *one row* that carries *counts* of how many providers receive it — not thousands of rows. Section 6 walks through exactly how, with example rows.

2 Glossary — every acronym and key term

2.1 Acronyms

Acronym	Meaning
BCBS / BCBSTX	Blue Cross and Blue Shield (of Texas). The insurer whose data this is. BCBSTX is operated by HCSC (Health Care Service Corporation), which also runs the Blue plans in IL, OK, MT, and NM.
MRF	Machine-Readable File. The large computer file (compressed JSON) an insurer publishes with its negotiated prices.
TiC	Transparency in Coverage. The federal rule requiring insurers to publish MRFs.
CPT	Current Procedural Terminology. The 5-digit code system for physician/professional services (e.g., `99214` = an established-patient office visit). Owned by the AMA.
HCPCS	Healthcare Common Proc; edure Coding System. Codes for items/services not in CPT (e.g., `G0121` = screening colonoscopy). CPT is technically "HCPCS Level I."
MS-DRG	Medicare Severity Diagnosis-Related Group. A code for a whole inpatient hospital *stay* (e.g., `470` = major joint replacement). Used for facility inpatient pricing.

[1] Texas 2036, "Health Care Price Transparency," https://texas2036.org/health-price-transparency/ — a project that scores Texas hospitals' compliance with hospital price-transparency rules and shows how prices for shoppable services (e.g., childbirth, knee arthroscopy) vary across the state. *That* work uses files published by *hospitals*; *this* project uses files published by an *insurer* (BCBS Texas). They are cousins under the same family of price-transparency rules.

Acronym	Meaning
NDC	National Drug Code. Identifies a specific drug product. Used in the prescription-drug data.
POS	Place of Service. A 2-digit code for where care happened (e.g., `11`=office, `21`=inpatient hospital, `22`=hospital outpatient, `23`=emergency room).
NPI	National Provider Identifier. A unique 10-digit ID for a provider ; either an individual clinician ("Type 1") or an organization such as a hospital ("Type 2").
TIN	Tax Identification Number. A provider group's tax ID. Can be an EIN or an NPI.
EIN	Employer Identification Number. A 9-digit IRS tax ID for an organization/employer.
NPPES	National Plan and Provider Enumeration System. The federal registry that maps every NPI to a name, type, and address. We use it to find which providers are in Texas.
ASO	Administrative Services Only. A "self-funded" arrangement: a large employer pays its own members' claims and just hires BCBSTX to administer the plan. Contrast with "fully-insured."
OON	Out-of-Network. Care from a provider not in the plan's network. (We did *not* include out-of-network "allowed amount" data in this run.)
PBM	Pharmacy Benefit Manager. The entity that negotiates drug prices; relevant to the prescription-drug file.
MPFS / OPPS / IPPS	The three Medicare payment systems we benchmark against: Physician Fee Schedule (professional), Outpatient PPS (hospital outpatient), Inpatient PPS (hospital inpatient). "PPS" = Prospective Payment System.
DSHS / PHR	Texas Dept. of State Health Services / Public Health Region. Texas is divided into 8 administrative public health regions; we use them for geographic price comparisons.

2.2 Key data terms (how we use them)

Provider

A person or organization that delivers care, identified by an **NPI**. "Provider" can mean an individual clinician (Type-1 NPI) *or* an organization like a hospital (Type-2 NPI).

Provider group

The insurer's unit for attaching a price: a bundle of one or more NPIs that share a tax ID (**TIN**) and get the same negotiated rate. In the source file it has an integer id (`provider_group_id`) that is unique *only within one source file*.

Organization / company

We approximate this by the **TIN** (tax ID) + organization name on a provider group. It is noisy ; insurers sometimes file an individual clinician as their own "organization." It is *not* a clean hospital-system identifier.

Network

A named set of providers a plan contracts with (e.g., "Blue Choice PPO," "Blue Essentials").

Prices are published per network. One insurer has many networks.

Payer / insurer
The company publishing the prices (here, BCBS Texas). A future run could add other payers (UnitedHealthcare, etc.).

Coverage segment
Whether a plan is **fully-insured** (the insurer bears the risk) or **self-funded / ASO** (a self-funded employer bears the risk, insurer just administers). A price is published per *network*, not per segment, so most prices serve *both* (see Section 6).

Snapshot
One monthly download/extraction, identified by the variable `run_date`. Stacking snapshots over time builds a price history (Section 6.7).

3 Where the data comes from

BCBS Texas publishes through two doors; we use both.

Route	Entry point	What it gives us
Fully-insured	Public page bcbstx.com/member/machine-readable-files links a ∼27 MB index listing ∼6,900 files.	All BCBSTX fully-insured plan networks. After Texas scoping: **33** unique in-network files.
Self-funded / ASO	`si-filelist.json` lists every self-funded employer (3,018 in Texas), each pointing to its own index.	The same shared national network files (the negotiated prices are shared), plus a few ASO-only networks. Texas total: **35** unique files.

Why this is manageable

The negotiated-price files are *shared nationally* — the same "Blue Choice PPO" file is referenced by thousands of employers. After keeping only Texas networks and removing duplicate file links, the entire Texas footprint is ∼35 files / ∼12 GB compressed, not the thousands of large files that the raw indexes suggest.

4 Scope decisions — what is and isn't included

Every step narrows the data. Here is each decision, in full sentences, with an example.

Decision	What it controls	Our choice this run (with an example)
Coverage segments	Which kinds of plans to pull.	**Both** fully-insured and self-funded. Example: we include both the individual "Blue Advantage HMO" plans and the self-funded employer plans that use "Blue Choice PPO."
Networks (geography of the contract)	Which provider networks to keep.	**Texas networks only.** We drop out-of-state "Blue Card" networks (a TX member who gets care in Tennessee). Example: we keep "Blue Choice PPO" and "Blue Essentials"; we drop "TN Blue Network P."
Procedures	Which billing codes to keep.	The policy team's menu of **93** codes across 11 categories (Section 7). Example: we keep CPT `5940; 0` (vaginal delivery) and MS-DRG `470` (joint replacement); we ignore the other ∼10,000 codes in the file.
"Texas provider"	Which providers count as in Texas.	A provider whose *practice-location state in NPPES is Texas*. We do *not* count a provider who only has a Texas *mailing* address. Example: a Houston clinic (practice ZIP 770xx) is kept; a Louisiana clinic that bills a Texas plan is dropped.
Unresolvable providers	Rows we cannot tie to a real Texas provider.	**Dropped, no guessing.** Two cases: (a) a price attached only to a tax ID with a placeholder NPI of `0`; (b) a provider whose NPPES state is not Texas. Both are removed.
Prescription drugs	Whether to include the separate drug file.	**Included**, as a separate drug-level dataset (different unit ; a drug, not a procedure).
Out-of-network amounts	Whether to include OON "allowed amounts."	**Not this run.** They are heavily suppressed (hidden when fewer than 20 claims), so they are sparse.

Decision	What it controls	Our choice this run (with an example)
Payment arrangements / bases	Bundles, capitation, percentage-of-charge, per-diem.	**Kept and flagged, not deleted.** Analysis restricts to comparable fee-for-service dollar prices (Section 5), but the others remain in the file for separate study.

To change the scope

Edit `config/target_codes.csv` (procedures) or `config/target_providers.csv` (specific providers) ; no code changes needed. Because the filter runs *during* extraction, adding procedures triggers a fresh data pull and full rebuild. Section 7.1 spells out exactly what re-runs and whether it changes the findings; Section 7 has step-by-step scenarios.

5 How the data was constructed (de-dup, roll-up, compression)

The raw files are tens-to-hundreds of GB uncompressed. Five techniques make them tractable; each is a deliberate, documented choice.

1. **Stream, never load.** We read the compressed file a little at a time (a "streaming" JSON parser) and never hold the whole thing in memory.
2. **De-duplicate file links.** The same network file is referenced by thousands of employers. We collect the list of file web-addresses and keep each *unique* one once. Example: 2 sample employers referenced 255 file links that pointed to the *same* handful of actual files.
3. **Filter early.** We drop any procedure not on the menu and any provider not in Texas *while streaming*, before anything is written.
4. **Collapse (roll-up) across providers.** The key move (detailed in Section 6): one row per distinct price, with provider *counts* instead of one row per provider.
5. **Slim + compress the provider directory.** The full provider list is reduced to one row per provider group (organization, tax ID, network, Texas-NPI count, region) and the very large raw version is stored gzip-compressed (`providers.csv.gz`, ∼76 MB instead of ∼900 MB).

5.1 Two data quirks you should understand (with examples)

Quirk 1 — "rate basis": not every number is a dollar amount

The variable `rate_basis` tells you *how to read* `negotiated_amount`:

- `negotiated` or `fee schedule` — a **dollar** price (e.g., $103.14). Comparable.
- `percentage` — a **percent of the provider's billed charge** (e.g., the value `40` means "40% of whatever the hospital bills"). There is *no* fixed dollar amount, so it *cannot* be compared to a dollar price.
- `per diem` — a **per-day** amount for an inpatient stay, not a total.

Consequence: mixing these inflates apparent price variation (a "40" sitting next to a "$15,000" is nonsense). So every price analysis filters to the comparable dollar rows using the flag `dollar_ffs == 1` (fee-for-service + dollar basis). About 82% of this run's rows are comparable dollars; the rest are kept but excluded from dollar comparisons.

Quirk 2 — "place-of-service bundling": one price covers many settings

In the BCBSTX *professional* file, a single negotiated price is attached to a whole *set* of place-of-service codes at once. A real example value of `place_of_service` is `11;17;20;50;72;81` — one office-visit rate that applies to the office, an urgent-care center, a school clinic, and more, *simultaneously*. **Consequence:** you *cannot* use this file to compare the price of the same service in an office (POS 11) vs. a hospital outpatient department (POS 22) — they share one rate. The feasible "site of care" comparison here is therefore *professional* (the clinician's fee) vs. *institutional* (the facility's fee), which *are* priced separately (Section 10).

5.2 Quirk 3 — "ghost rates" (the key methodological limitation)

What they are. Federal transparency requirements mandate that insurers publish a negotiated price for *every* provider-and-service combination within their contracts. Consequently, these files are heavily populated with rates for clinical combinations that never actually occur in practice — for example, a pediatric clinic possessing a published rate for a hip replacement. These token, placeholder figures are formally termed **ghost codes** or **ghost rates** (Pathak & Muhlestein, 2024). They are legitimate contract terms rather than administrative errors, but they represent services for which no patient is ever billed or treated. Recent analyses demonstrate that between **70% and 92%** of all prices in these files are ghost rates. Because the files inherently lack the claims data needed to indicate how often each service is rendered, these placeholder rates cannot be precisely filtered out from the files alone.

Why they distort the numbers. The overwhelming prevalence of low-value, unutilized ghost rates systematically drags median and mean values downward, artificially indicating that published prices are lower than amounts actually paid. For example, simple medians of BCBS Texas *professional* rates put an office visit (CPT 99214) at ≈63% of Medicare and an MRI read at ≈75% — yet commercial insurers typically reimburse physicians *above* Medicare. This apparent discount is largely a ghost-rate artifact: the professional files are saturated with never-used listings that skew the median and understate what BCBS Texas actually pays for realized visits. *Facility* rates — e.g., a knee-replacement facility fee at ≈100% of Medicare, or the joint-replacement DRG at ≈148% — are much less affected and behave as expected (commercial above Medicare). Section 8 works this finding through in detail — including why a simple "below 25% of Medicare" filter does *not* dispose of it. A concrete Texas illustration: filtering analyses found a lumbar discectomy's

volume-weighted median of $8,250 collapsed to a ghost-weighted median of just $1,500 (Texas Hospital Association / Denniston Data, 2024).

How we locate and handle ghost rates.

- **The gold standard** (Pathak & Muhlestein, 2024) cross-references each provider's NPPES *taxonomy* (specialty) against the specialty a billing code implies, and flags clinically implausible pairs (a psychiatrist's rate for a heart procedure) — often using external claims to conf; irm a pair has zero real volume. The 2026 federal proposed rule would *mandate* this taxonomy cross-reference to remove clinically implausible rates, estimating it alone shrinks the files by ≥70%.
- **What this project does today.** Our analytic price table is collapsed across providers, so it does not carry per-rate taxonomy ; we therefore do *not* yet remove ghosts row-by-row. Instead `600_analysis.do` adds a transparent *heuristic locator*: where a Medicare benchmark exists, it flags prices below ≈25% of Medicare as "implausibly low / likely ghost," reports how many there are, and shows the median *with vs. without* them so the distortion is visible. This is a floor-based heuristic, not the taxonomy gold standard.
- **Three interim principles** we follow until external volume data is integrated: (1) report the *median* rather than the mean (robust to the skew); (2) restrict to comparable dollar-based fee-for-service rates (exclude percentage-of-charge and per-diem); (3) describe metrics strictly as "prices *published*" rather than "prices *paid*."

The real fix — volume weighting. To reflect true market pricing, analyses should incorporate external *volume* (how often each service is billed) to compute a volume-weighted median, which inherently down-weights unused ghost listings. For Texas, the natural source is the **Texas All-Payer Claims Database (TX-APCD)** (which ties into Texas 2036's broader APCD work); alternatives include Medicare service counts (CMS Physician/Supplier Procedure Summary) or commercial-claims datasets such as MarketScan or the Health Care Cost Institute (for a fee). See Section 10.4.

References.

- Pathak, Y., & Muhlestein, D. (2024). *Hospital system market share and commercial prices: a cross-sectional approach using price transparency data.* Health Economics Review, 14. https://doi.org/10.1186/s13561-024-00580-w — and the companion analysis of 61 insurers (incl. 58 BCBS plans) finding 91.8% of listed rates are ghost rates, detailing the NPI/claims method to identify them.
- *Over- and underreporting of prices: most hospitals are not compliant with the Hospital Price Transparency Rule* (PMC, 2024) — finds 12% of hospitals report prices for services they do not provide (the hospital-side parallel). https://pmc.ncbi.nlm.nih.gov/articles/PMC11363865/
- Health Affairs Forefront (Mar. 2026). *Taking Stock of Proposed Updates to Health Plan Price Transparency Rules* — the proposed HHS/Labor/Treasury rule mandating removal of ghost rates via provider-taxonomy cross-reference (≥70% file-size reduction).
- Texas Hospital Association / Denniston Data (May 2024). *Making Dollars & Sense from Transparency in Coverage MRFs* — ties ghost rates to the No Surprises Act / Qualifying Payment Amount dispute; the $8,250 → $1,500 lumbar-discectomy example. https://www.tha.org/blog/making-dollars-sense-from-transparency-in-coverage-mrfs/

- Peterson Center on Healthcare (2024). *Insights from a Discussion of Transparency in Coverage Data* — researcher consensus on filtering ghost codes and reducing the 600 GB+ files. https://petersonhealthcare.org/news/insights-from-a-discussion-of-transparency-in-coverage-data/

6 The unit of analysis — what one record holds, and how to use it

"Unit of analysis" is jargon for one simple question: *when you open the dataset, what does a single row stand for?* Here the answer is **one row = one distinct negotiated price for one procedure in one provider network**. It is *not* one row per patient, per claim, or per provider. Getting this straight is the difference between a correct statistic and a meaningless one — so this section spells out what each field on a record holds, shows real rows, and walks through how those rows become the analyses elsewhere in this guide.

6.1 What each field on a record holds

A record answers: *who* (network) pays *what* (amount, and how to read it) for *which service* (code), to roughly *how many* providers (counts), *where* (region). The core fields:

Field on the record	What it means	Example
`procedure_code` (+ `code_system`)	The billing code for the service, and which code system it is from.	`99214` (CPT)
`service_label`	Our friendly name for that code (from the menu).	Office visit est pt
`provider_type`	Whose fee this is: the clinician's (*professional*) or the facility's (*institutional*). Never mix the two.	professional
`network_name`	The provider network the price applies to.	Blue Choice PPO
`region`	The Texas Public Health Region of the providers carrying this price.	2/3
`rate_basis`	**How to read the amount**: a dollar price (`negotiated` / `fee schedule`), a percent of charge (`percentage`), or a per-day amount (`per diem`).	negotiated
`place_of_service`	The *set* of care settings this one price covers (one price often spans many).	`11;17;20;...`
`negotiated_amount`	The price itself ; a dollar figure when `rate_basis` is a dollar basis.	93.89
`n_provider_groups`	How many provider groups are paid this exact price.	49
`n_tax_ids`	How many distinct organizations (tax IDs) are paid it.	49

Field on the record	What it means	Example
`n_npi_memberships`	A provider head-count behind the price (double-counts shared NPIs ; read as "breadth," not a clean count).	53

Two siblings round out the picture: `run_date` (which monthly snapshot ; Section 6.7) and `market_segment` (fully-insured / self-funded / both).

6.2 From the raw file to one record (the collapse)

Inside the raw file a single procedure carries a list of rates, and each rate is attached to a (often huge) list of provider groups. Flattened naively that is one row per *procedure* × *provider-group* × *price* — about **2.06 million** rows for one mid-size network, with a matching **37.5 million**-row provider list; across all 35 Texas files, roughly a *billion* rows, mostly the *same price repeated*. We **collapse** to the **distinct price**: identical prices become *one* record carrying *counts* of how many providers receive it.

Worked example: the collapse

BCBS Texas pays $103.14 for an office visit (CPT 99214) to 49 provider groups in Dallas–Fort Worth. The raw file implies 49 rows; we store **one** record: `99214`, $103.14, region `2/3`, `n_provider_groups`=49. A group paid $81.95 is a *second* record. One mid-size network went from **2,055,938** raw rows to **18,414** distinct-price records — no price information lost.

6.3 Sample records — what the data actually looks like

Real rows (lightly simplified). Watch the variety the fields capture: a single office-visit price that spans a *set* of place-of-service codes; a facility DRG record; and a `percentage` row whose "95" is *not* $95 but **95% of the hospital's billed charge** (why we exclude such rows from dollar comparisons).

network	reg.	code	prov. type	rate_basis	amount	POS	n_grp
Blue Essentials	7	99214	professional	negotiated	$93.89	11;17;20;...	1
Blue Choice PPO	2/3	99214	professional	negotiated	$103.14	11;17;...	49
Blue Choice PPO	4/5N	470	institutional	negotiated	$12,571	(DRG)	1
Blue Choice PPO	6/5S	99395	institutional	**percentage**	**95**	(NULL)	1

The other three tables carry their own units: `bcbstx_rx_prices.dta` (one row = one distinct *drug* price, by NDC), `bcbstx_provider_org_directory.dta` (one row = one organization in one network), and `bcbstx_medicare_benchmark.dta` (one row = one code's Medicare allowed amount).

6.4 How records become analysis (conceptually)

Every headline result is a simple operation *over these records* — the trick is to filter to comparable records first (Section 10) and respect what a row is:

- **Price variation for a service** (the dispersion ranking): take *every* record for one code + provider type (e.g. all professional `99214` dollar rows), find the 10th and 90th percentile of `negotiated_amount`, and divide (p90/p10). Each record is one distinct price; `n_provider_groups` tells you how much of the market sits behind each.
- **Median price by region**: keep one code, group records by `region`, take the median `negotiated_amount` in each — exactly the regional map and table.
- **Percent of Medicare**: join each record to `bcbstx_medicare_benchmark.dta` on the code and divide `negotiated_amount` by the Medicare amount.
- **Professional vs. facility**: `provider_type` is on every record, so split a code's records by it and compare medians — the feasible "site of care" lens.
- **Cross-network / segment**: group by `network_name` or `market_segment`.

The one rule that keeps the math honest

A record is a *price*, not a patient. These records measure *price levels and variation*, never *spending* — that needs how *often* each service happens (volume), which the files do not contain (Section 10.4). And only *comparable* records belong in one number: keep dollar `rate_basis` rows (a "95" percentage next to a "$15,000" is nonsense) and never pool professional with institutional.

6.5 What got "collapsed away" — and where it went

- **Individual provider identities** are not on the price record (the 37-million-row part); they live in the separate **provider directory** (`bcbstx_provider_org_directory.dta`). To attach named providers to prices, re-run extraction with the `--emit-group` switch.
- **Provider geography** is summarized to the *modal* (most common) county/region of the groups carrying a price (Section 9); the underlying ZIPs are in the NPI crosswalk.

6.6 What `--emit-group` would unlock: named providers behind every price

As of now, the analytic price table is *collapsed across providers* (Section 6): one row is one distinct price, carrying only *counts* of how many provider groups receive it. The individual identities — which specific clinics, hospitals, and systems are paid each rate — are deliberately set aside (that is the + 37-million-row part). Re-running the extractor with `--emit-group` keeps that link: it writes a **rates-by-group** bridge table with *one row per price* × *provider group*, so each negotiated rate can be traced to the named organizations (and, through their NPIs, the systems and clinicians) that carry it.

What the file would look like. Where the $103.14 office-visit rate is *one* collapsed row today (with `n_provider_groups`=49), the rates-by-group file would hold *49* rows — the same price repeated once per group — each carrying a `provider_group_id` that joins to the per-group provider crosswalk for a tax ID, organization name, NPIs, and location:

network	proc.	amount	group id	tax id	organization (from the crosswalk)
Blue Choice PPO	99214	$103.14	4471	75…12	Acme Family Medicine PA (Dallas, 48113)
Blue Choice PPO	99214	$103.14	4472	74…90	Lakeview Internal Medicine (Collin, 48085)
Blue Choice PPO	99214	$103.14	…	…	… (47 more groups at this same rate)
Blue Choice PPO	99214	$81.95	5012	76…31	Rio Grande Clinic (Hidalgo, 48215)

How much bigger it gets. The collapse shrank one mid-size network from ≈2.06 million raw rows to 18,414 distinct prices — about a **110**× reduction. Undoing it at the group level is roughly that factor in reverse: the rates-by-group file is the *sum of* `n_provider_groups` across today's ≈2.1 million price rows, plausibly **tens to a few hundred million rows** statewide. The finer one-row-per-NPI grain (the `n_npi_memberships` level) is larger still — approaching the ≈1 billion rows this guide cites for the raw provider list. In storage that is a jump from today's ≈0.7 GB `prices.csv` to a **multi-gigabyte to tens-of-gigabytes** artifact (a few GB gzipped). **Practical implication:** you would almost never emit the whole thing — you would scope `--emit-group` to a handful of codes or one region (Section 7.1), yielding a few million rows a single machine handles comfortably.

How you would work with it (a different, relational merge). The extant table is self-contained. The rates-by-group file is instead a *fact* table you join to the per-group provider crosswalk (the slim `providers.csv.gz` the extraction already writes) and, in turn, to an external NPI→hospital-system (**CCN**) crosswalk to name systems:

```
use "rates_by_group.dta", clear                 // 1 row per price x provider group
merge m:1 source_file provider_group_id ///     // group id is unique only WITHIN a file
      using "providers.dta"                     // -> tax id, org name, NPIs, county/region
* then join the tax id / NPI to an NPI -> hospital-system (CCN) crosswalk to name systems
```

Two cautions: `provider_group_id` is unique only *within* a source file, so the join key is `(source_file, provider_group_id)` and any *cross-file* provider analysis must first resolve groups to a stable identity (tax ID or NPI); and a provider recurs across thousands of prices, so de-duplicate deliberately before counting providers. **Also merging these much larger files across time or across insurance companies would be likely infeasible.**

What this would let us answer (the payoff)

With provider identity restored and much larger files, we move from "how much do prices vary?" to "*who* is expensive, and *why*":

1. **Name the high- and low-price providers and systems** for any service — e.g. "System A's median knee replacement runs well above the regional median," by name rather than in aggregate.
2. **Provider-level, within-market variation:** for one service in one county, how far apart are *named* providers — the true "shop here, not there" picture.
3. **Market concentration / negotiating leverage:** how few systems set most of a market's prices (with provider counts now, and a proper **HHI** once volume is added; Section 10.4).
4. **The real ghost-rate filter (the big one):** attach each provider's NPPES *taxonomy* (specialty) and flag rates for services outside that specialty — the gold-standard, taxonomy-based identification the crude "below 25% of Medicare" floor cannot do (Sections 5.2 and 8). This is the path to a *paid*-like % of Medicare.
5. **Exact geography:** use each provider's own location instead of the group's modal region — enabling a true (not approximate) price map by county, ZIP, and legislative district (Section 9).
6. **Network design:** which providers sit in narrow vs. broad networks, and whether narrower networks buy lower rates.
7. **Payer × provider leverage:** combined with other payers (Section 13), "does System A get a higher rate from BCBS than from UnitedHealthcare for the same service?"
8. **Access and equity overlays:** join named provider locations to demographics, shortage-area (HPSA) designations, or hospital quality ratings for an access or "value" lens.

What it still would *not* give us

Provider identity is not *volume*: rates-by-group still says nothing about how often each provider performs a service, so *spending* and true market *shares* stay out of reach without an external claims source (Section 10.4). And organization names are filed by the insurer and are noisy (individual clinicians sometimes appear as their own "organization"), so naming *systems* reliably needs the external NPI/TIN→system crosswalk — not the raw `provider_org_name`; and I'm not sure we can get this information in a meaningful way.

6.7 How to maintain: Do I have to merge or append anything going forward?

- **Within one monthly run: no.** The pipeline already stacks every network and both coverage segments into one `bcbstx_prices.dta` — `network_name` and `market_segment` are columns, not separate files.
- **Across months: append, then it "just works."** Each run is one **snapshot** stamped with `run_date`. `append` a new month onto the prior file; the analysis do-files key on `run_date`, so cross-sectional tables use the *latest* snapshot and time-trend panels switch on once two or more snapshots exist.

```
use "02_cleaned/bcbstx_prices.dta", clear         // this month
append using "_archive/2026-04/bcbstx_prices.dta" // last month
```

Fully-insured vs. self-funded: a subtlety

A negotiated price is published *per network*, not per plan. The same "Blue Choice PPO" price serves both fully-insured and self-funded members, so `market_segment` is `both` for shared networks. The segment distinction lives in *which plans point to a network*, not in the price.

7 The scope-control files (how to change what gets counted)

Three small, human-editable files in `_code/_pipeline/config/` control exactly what the download and extraction keep. **Editing them is how you change scope — you do not touch any code.**

File	What it does
`target_codes.csv`	The procedure menu. One row per billing code, with columns `billing_code_type` (CPT/HCPCS/MS-DRG), `billing_code`, `short_label`, `category`. **Only codes listed here are extracted**; everything else in the insurer file is skipped. It is also the join that gives each price a friendly `service_label` and `service_category`.
`target_providers.csv`	An *optional* provider filter. If it has only a header (no data rows), all Texas providers are kept. Add NPIs/TINs here to restrict to specific systems.
`npi_tx_xwalk.csv`	(Built automatically by stage 0, not hand-edited.) The list of Texas NPIs with practice ZIP and Public Health Region. This is what enforces "Texas providers only" and powers the region analysis.

Scenario: "Add MRI of the knee and three behavioral-health visits"

1. Open `config/target_codes.csv`. 2. Add rows, e.g.
`CPT,73721,MRI lower extremity joint wo contrast,Imaging`
`CPT,90834,Psychotherapy 45 min,Behavioral health`
3. Save. 4. Re-run the pipeline (the download step re-pulls and re-filters; Section 11). The new codes will appear in `bcbstx_prices.dta` and in every table/figure automatically. **Note:** adding codes means re-downloading, because the filter happens during extraction — the codes you didn't request were never written to disk.

The version the policy team reviews (the full menu with rationale and a blank "include?" column) is `_documentation/data_elements_menu.csv`.

7.1 Adding codes later: what re-runs, and does our menu bias the findings?

The policy team will likely want to add services over time. Two honest questions come with that.

1. How do I add a service, and what happens next? Add one row per new code to `config/target_codes.csv` (columns `billing_code_type`, `billing_code`, `short_label`, `category`); to add a new *data field* rather than a new code, extend the column list in `02_extract.py`. Then **a full pull-and-rebuild is required**, in this order:

1. In Stata, set `global force_download "1"` and run `100_data_download.do`. This is the slow step. The extractor keeps *only* menu codes *while streaming*, and the raw files are never stored on disk — so a new code's prices were never written anywhere. The only way to get them is to re-download and re-stream the ∼12 GB of source files.
2. Run `200`–`600` (or `_runall.do`) to rebuild the `.dta` files, every table and figure, and the website; then recompile this guide.

There is no shortcut that avoids the re-download — that is the price of keeping the stored data small. In practice, fold new codes into the *next monthly pull* rather than re-pulling mid-month.

2. Does adding codes change the existing results — i.e. are we biased by the original menu? Separate two ideas:

- **Within each selected service: no bias, and adding codes changes nothing.** For every one of the 93 codes we already keep *all* Texas-provider negotiated prices — every network, both coverage segments, every region. We do not sample. So the statewide and regional patterns, the price dispersion, and the % of Medicare for those services rest on the *full population* of BCBS Texas prices for them (subject only to the documented ghost-rate and county-resolution caveats). Adding a 94th code adds a new service to the analysis; it does *not* move any number for the 93 already there.
- **Across services: the menu is a deliberate sample, so frame claims to its scope.** The 93 codes are a *purposive* selection of policy-relevant, shoppable, high-volume services (the policy team's choice), not a random draw from the ∼10,000 codes in the file. We can say "for common shoppable services, negotiated prices vary X-fold across Texas," but not "the average price of *all* care" or "total market prices" — those need the full code set *and* the volume data we do not have (Section 10.4). By design the menu leans toward shoppable outpatient procedures, imaging, maternity, and common inpatient stays: the right lens for the shopping-and-variation story, the wrong lens for a whole-market average.

Bottom line on the menu

Our findings are *scope-limited, not biased.* For the services we chose, the statewide and regional claims are complete and sound. Adding codes *widens* the story (more services covered); it does not *correct* anything already reported. A true "all of health care" or market-average claim is a separate, larger project (full code set + a volume / claims source).

8 Percent of Medicare — how it was built, and a headline finding

A natural question for any negotiated price is "how does it compare with what Medicare pays?" Dividing each commercial price by the Medicare allowed amount for the same service gives one interpretable yardstick ("100% = the same as Medicare"), the most-cited benchmark in this field (e.g. the RAND Hospital Price Transparency Study). This section documents exactly how we built that benchmark, what we found, and — because the finding is striking — why a tempting one-line "ghost-rate filter" does *not* dispose of it.

8.1 How we built the benchmark (data sources and formula)

A separate stage, `build_medicare_benchmark.py`, downloads three CMS fee schedules and reduces each to one number per code — the *national* Medicare allowed amount — writing them to `bcbstx_medicare_benchmark.dta` (columns `procedure_code`, `code_system`, `fee_schedule`, `applies_to`, `medicare_allowed`). Which schedule applies depends on how the service is billed:

Billed as	CMS source dataset	How we turn it into one dollar amount (the column `medicare_allowed`)
Professional (clinician fee)	Physician Fee Schedule (**MPFS**; the `PPRRVU2026` file in CMS release `RVU26A`)	(Work RVU + non-facility Practice-Expense RVU + Malpractice RVU) × the CY2026 conversion factor \$33.4009. Active-status codes only; global service (no `26`/`TC` modifier).
Outpatient facility	Hospital Outpatient PPS (**OPPS** Addendum B, Jan. 2026)	the national "Payment Rate" column listed for that HCPCS code.
Inpatient stay	Inpatient PPS (**IPPS** Final-Rule Table 5)	the MS-DRG relative weight × the FY2026 standardized base of \$7,276.76 (\$6,752.61 operating + \$524.15 capital).
Clinical lab	Clinical Lab Fee Schedule (**CLFS**)	*Not built this run* (source URL unconfirmed); the 8 lab codes therefore carry no benchmark.

Then, in `600_analysis.do`, every price is tagged with which schedule applies (the `applies_to` field: *professional*, *institutional inpatient*, or *institutional outpatient*, derived from its provider type, care setting, and code system) and joined to the benchmark on the procedure code plus that tag. The ratio is then

$$\texttt{pct\_of\_medicare} = 100 \times \texttt{negotiated\_amount} \,/\, \texttt{medicare\_allowed}.$$

All amounts are *national* (every geographic index set to 1.0 — a first-pass simplification). Of the ≈1.73 million comparable dollar prices this run, ≈1.44 million (≈83%) matched a benchmark.

Worked example (CPT 99214, an office visit)

Medicare's MPFS allowed amount is (Work + Practice-Expense + Malpractice RVUs) × \$33.4009 = **\$135.61**. BCBS Texas's median negotiated 99214 is **\$85.38**. So 100 × 85.38/135.61 = **63% of Medicare**. (A code can have more than one benchmark: 27447 carries a *professional* MPFS amount of \$1,159 *and* an outpatient-facility OPPS amount of \$13,117; `applies_to` selects the right one for each price.)

8.2 The headline finding

The pattern is consistent and, at first glance, surprising (medians of comparable dollar prices, this run):

Service	% of Medicare	Reading
Office visit, established 30–39 min (CPT 99214, prof.)	≈63%	below Medicare
Office visit, established 20–29 min (CPT 99213, prof.)	≈59%	below
Diagnostic colonoscopy (45378, prof.)	≈71%	below
MRI lumbar spine (72148, prof.)	≈75%	below
All professional codes (median)	**≈75%**	*below*
Cataract surgery with IOL (66984, prof.)	≈106%	above
Knee replacement, facility (27447, OPPS)	≈100%	at par
All institutional codes (median)	**≈108%**	*above*
Joint-replacement inpatient stay (MS-DRG 470, IPPS)	≈148%	well above

So BCBS Texas's published **professional** (clinician) rates sit *below* Medicare — a statewide median around **75%** — while **facility** and **inpatient** rates sit *above* (median ≈108%, up to ≈148% for major stays). The facility/inpatient side behaves exactly as health economics expects: commercial plans pay hospitals well above Medicare. The professional side does *not*: commercial insurers almost universally pay *physicians* above Medicare, so a 75% figure is implausible as a description of what is actually *paid*. That tension is the finding — and it points straight at ghost rates.

8.3 Why a simple "below 25% of Medicare" filter does not fix it

A reasonable first instinct (and one we implement as a transparency check in `600_analysis.do`, "section 5b") is: *flag prices below 25% of Medicare as implausibly low / likely ghost, and see how much they move the result.* Here is exactly what happens before and after, for professional prices:

- Of the ≈1.27 million professional prices with a benchmark, only **4.4%** (≈55,000) fall below 25% of Medicare.
- Dropping them moves the professional median only from **75.3%** to **76.9%** of Medicare — a **1.6-percentage-point** shift.

The cheap floor barely budges the answer: even after removing the obviously-implausible bottom, professional rates still sit ≈77% of Medicare. **The simple filter cannot rescue the number, because ghost rates are not concentrated at the bottom.** A placeholder rate for a service a provider never performs can be set anywhere in the distribution — 30%, 60%, even above 100% of Medicare — so no single low threshold separates ghosts from real rates. The peer-reviewed estimate is that ≈92% of listed rates are ghosts (Pathak & Muhlestein, 2024; Section 5.2); a 25% floor catches only the most extreme sliver of them.

Why we flag, but do not hard-drop, ghost rates

A "below 25% of Medicare" cut is a crude proxy, not a clean filter. Hard-dropping on it would (1) wrongly delete *legitimate* low prices; (2) still miss the majority of ghosts, which sit higher in the distribution; (3) only work for the ≈83% of rows that match a Medicare benchmark at all; and (4) impose false precision. The defensible move is the one we take: report the median *with vs. without* the flag so the reader sees the drag at the floor is small — which is itself the evidence that the real fix lies elsewhere. Identifying ghosts properly needs the provider's *specialty* (NPPES taxonomy vs. the specialty a code implies) and/or external *claim volume* — neither of which survives on our collapsed price rows.

8.4 What we would run with better data (and what each would change)

1. **Volume weighting (highest value).** Weight each price by how often the service is actually billed — from the Texas All-Payer Claims Database (TX-APCD), Medicare service counts (CMS Physician/Supplier Procedure Summary), or commercial claims (MarketScan, HCCI). Rarely-used ghosts get near-zero weight, so the professional ratio should rise toward — and likely above — 100%, turning a *published*-price figure into a *paid*-like one. This also unlocks spending and market-share measures (Section 10.4).
2. **Taxonomy filtering (the gold standard; Pathak & Muhlestein).** Cross-reference each provider's NPPES specialty against the specialty a code implies, and drop clinically implausible pairs. Requires a re-extraction that keeps per-rate provider taxonomy (our collapse drops it).
3. **Locality adjustment.** Replace the national amounts with Texas-specific ones — the **GPCI** (Geographic Practice Cost Index) for professional fees and the hospital wage index for facility — now feasible because every price carries a county.

Until at least the first is in place, every % of Medicare figure here describes prices *published*, not prices *paid* — and it should be stated that way in any external use.

9 Texas geography: county, Public Health Region, and ma; ps

We tag each price with the *county* of the provider group's practice location (the chain: NPPES practice ZIP → county, via the free Census ZIP-to-county file), and roll county up to one of the 8 DSHS Public Health Regions. *County* is the atomic geography; region is derived from it. About 99% of Texas NPIs were tagged. Because the price table is collapsed across providers, a price takes the *modal* (most common) county of the provider groups that carry it. **This is why the geography add-on required a fresh re-extraction:** the old collapsed prices had no link back to providers, so there was nothing to attach a county to.

Maps. `600_analysis.do` draws choropleth (color-shaded) maps with `geoplot` (with `spmap` kept as an automatic fallback) on Census shapefiles stored in `03_mapping/`. The set: a *county* map and a *Public Health Region* map of the median office-visit price (CPT 99214); an *approximate* office-visit price map by State Senate and House district (see below); and two provider-*supply* maps (by Senate district and by ZIP code) showing where providers concentrate. This mirrors the mapping pattern in the team's `NCES_EDGE` project.

First results (this run): of the ≈2.1 million distinct-price rows (2,097,564), ≈99% carry a county and region (the rest are providers registered out of state or with an unmapped ZIP). Geographic variation is real but modest for shoppable professional services ; e.g., the median professional colonoscopy (CPT 45378) ranges from ≈$234 in Region 11 (the lower Rio Grande Valley) to ≈$273

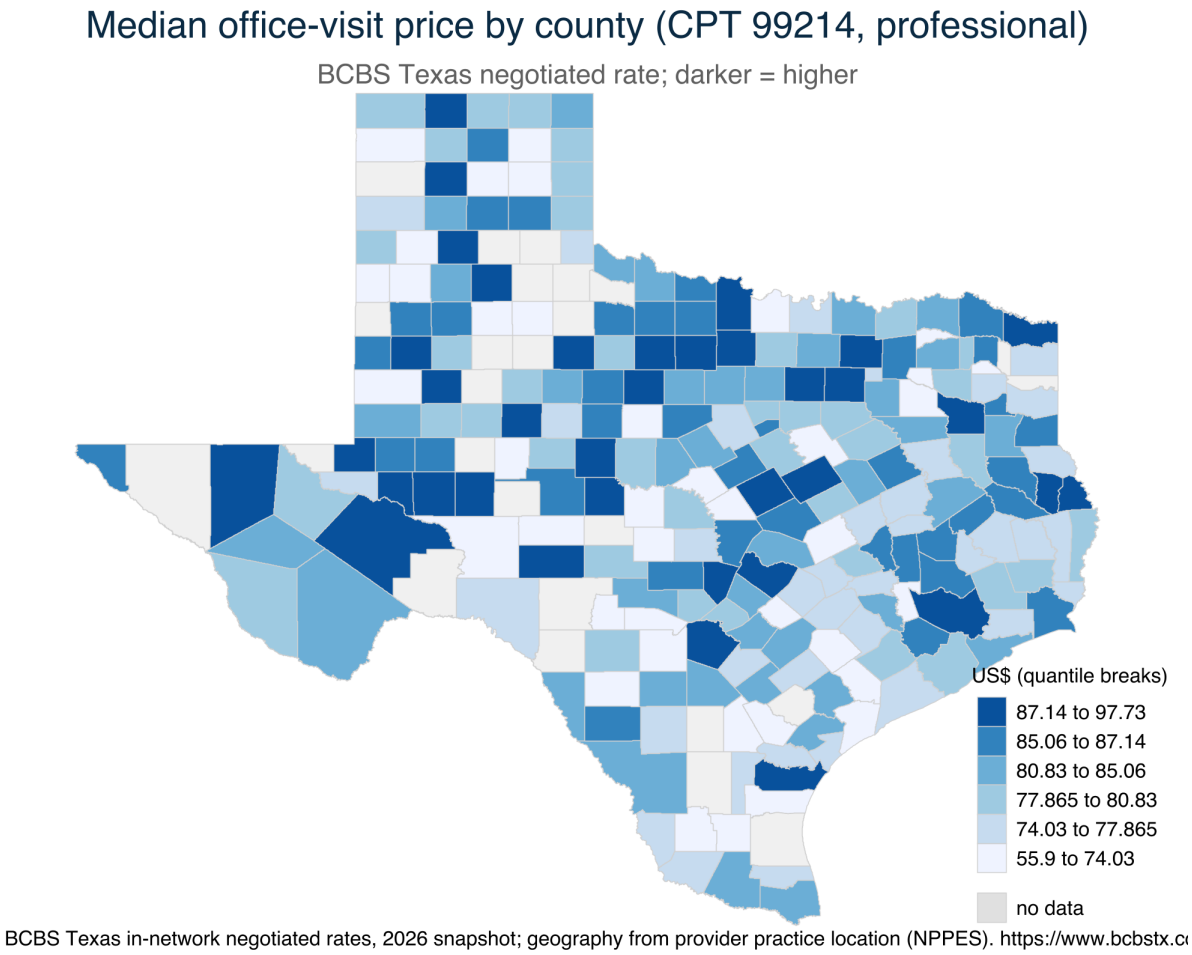

Figure 1: default

in Region 1 (the Panhandle / West Texas).

Legislative districts (geocoding done; supply and approximate price both mapped). The Texas State Senate and House shapefiles are staged in `03_mapping/` (`tx_sldu` = 31 Senate districts, `tx_sldl` = 150 House districts), and we have *completed* the point-in-polygon geocoding: each provider ZIP centroid (Census 2020 ZCTA Gazetteer) is assigned to the district whose polygon contains it, tagging ≈585,000 Texas NPIs with a Senate and House district (the `senate_district` and `house_district` columns in the NPI crosswalk). The analysis website (the "600 – Analysis" page) draws from this both a provider-*supply* map (by Senate district) and an *approximate* office-visit *price* map (by Senate *and* House district). The price map is labelled approximate on purpose: because prices are published at the *county* level and districts cross county lines, we assign each provider its c; ounty median and average within district. An *exact* price-by-district map would still need a district-keyed re-extraction of the raw files (which are streamed and not retained) ; so it reads broad patterns, not precise district figures.

ZIP vs. county. We map at the ZIP level only for provider *supply* (the analysis page draws a ZIP point map ; one dot per ZIP, shaded by provider count), never for price: *county* is the atomic price geography (the collapse buckets prices by county), so there is no ZIP-level price to plot without re-extracting keyed on ZIP.

One more open item ; locality-adjusted Medicare. The % of Medicare (Section 8) uses *national* fee-schedule amounts (all geographic indices set to 1.0); adding the Texas **GPCI** (Geographic Practice Cost Index ; Medicare's local cost adjuster for physician fees) and hospital wage-index (facility) adjustments — now feasible because each price carries a county — would make the benchmark Texas-specific. The price-by-district map and the locality adjustment both appear in the feasibility matrix.

10 Analyzing the data in Stata — work plans

The numbered do-files in `_code/` both *run* the analysis and *document* it (each is also a web page; Section 11).

10.1 First, the comparable-price filter

Every dollar analysis starts the same way:

```
use "02_cleaned/bcbstx_prices.dta", clear
keep if dollar_ffs==1                        // fee-for-service, dollar-denominated only
drop if missing(negotiated_amount) | negotiated_amount<=0
* never pool professional + institutional -- they are different markets
```

10.2 What `400_data_profiler.do` already produces

- Snapshot scope (how many prices, networks, procedures; how many monthly snapshots).
- Service-menu coverage — which of the 93 menu codes appeared, and an explicit list of any that did *not*.
- Rate-basis composition (the share of dollar vs. percentage vs. per-diem rows) — the data-quality audit.
- Professional-vs-institutional and care-setting counts; the place-of-service-bundling check.
- Sanity checks on the dollar amounts (zeros, implausible highs).
- Median price by service category; provider-geography (region) coverage.

10.3 What `600_analysis.do` already produces

1. **Dispersion ranking** — which services vary most (90th / 10th-percentile price ratio).
2. **Dispersion by region** — for the most-variable services, *where* they vary: median price across the 8 Public Health Regions, panel by panel.
3. **Interactive price table** — a searchable, sortable web table of the per-service price summary, built with the team's `statashiny` and embedded on the analysis page.
4. **Shoppable case studies** — price distributions (median with a 10th–90th-percentile range) for childbirth, knee arthroscopy, colonoscopy, MRI, cataract, joint/hip replacement.
5. **Professional vs. facility** price and variation.
6. **Cross-network and fully-insured-vs-self-funded** comparison.
7. **Percent of Medicare** (Section 8).
8. **Maps** — median office-visit price by county and Public Health Region; an *approximate* price map by State Senate and House district; and provider-supply maps by Senate district and ZIP (Section 9).
9. **Prescription drugs** — brand vs. generic, dispensing fees, same-drug spread.
10. **Time trend / convergence** — activates automatically with $\geq$2 monthly snapshots.

10.4 What you cannot do with this data (and why)

"Prices published, not spending" — what that limitation means

The files contain a price for each procedure-provider combination, but **not how often that service is actually used** (no volume, no claim counts). Because **spending = price $\times$ quantity**, with no quantity we can describe *price levels and variation* but cannot compute total *spending*, a market-average price actually paid, or budget impact. Concretely: a $50,000 price on a procedure nobody gets matters far less than a $200 price on something done a million times ; and this data can't tell those apart.

Are there better ways? Yes ; by bringing in an *external* volume source and weighting:

- Weight each price by Medicare service counts (CMS Physician/Supplier Procedure Summary) or by commercial claim frequencies (MarketScan, HCCI, or a state all-payer claims database), then report a volume-weighted average and compare it to the unweighted one.
- Validate against a known claims-based figure for one service (an "alignment check").

Until then, we use price-only language: "negotiated price" and "price variation," never "spending" or "cost." Other things still out of reach: cross-payer comparison (only BCBSTX so far), market concentration (the **Herfindahl–Hirschman Index**, HHI ; a standard measure of how few firms dominate a market; needs clean hospital identity *and* volume), and a price-vs-quality "value" view (needs CMS quality ratings joined).

10.5 Feasibility matrix assessing what's possible or next

`600_analysis.do` writes `03_output/feasibility_matrix.csv`: a subjective matrix of, for each standard price-transparency analysis, whether we can do it with the data we have. Every row is one of three statuses:

- **Feasible now** ; we run it with this dataset (price dispersion, professional-vs-facility, cross-network, % of Medicare, the county and region maps, Rx).

- **Needs addition** ; not possible today, but ONE specific, named add-on unlocks it (the matrix names which). Examples: cross-payer comparison needs other insurers run through the same pipeline; an *exact* price-by-legislative-district map needs a district-keyed re-extraction (an *approximate* one ; each provider's county price mapped to its district ; is already produced this run, and the geocoding is done); a lab % of Medicare needs the Clinical Lab Fee Schedule file.
- **Not feasible** ; cannot be done from insurer negotiated-rate files at all, no matter what we add, because the files lack the needed information. Examples: total *spending* and *market-concentration* measures both need *utilization* (how often each service happens), which these files simply do not contain.

The last column of the matrix gives the specific next step or caveat for each row.

11 Folder map and how to run it

The project mixes a **Python** half (download + extract the giant files) and a **Stata** half (clean + analyze). Here is how they connect (via a simple `-tree-` command):

```
BCBSTX/
|-- Open_Documentation.html      <- double-click: opens the website
|-- _code/
|   |-- 000_controlfile.do        Stata: paths + settings (run first)
|   |-- 050_preferences.do        Stata: brand colors / graph style
|   |-- 100_data_download.do      Stata -> calls the Python pipeline below
|   |-- 200_data_management.do    Stata: CSV -> labeled .dta
|   |-- 300_labels.do             Stata: plain-English variable labels
|   |-- 400_data_profiler.do      Stata: coverage + data-quality
|   |-- 500_aggregation.do        Stata: price-summary tables
|   |-- 600_analysis.do           Stata: the analyses + figures
|   `-- _pipeline/                THE PYTHON HALF
|       |-- scripts/
|       |   |-- 00a_fetch_nppes.py + 00_build_tx_npi.py   (Texas NPI list)
|       |   |-- build_geo_xwalk.py        (adds Public Health Region)
|       |   |-- build_medicare_benchmark.py (code -> Medicare $)
|       |   |-- 01_discover.py  02_extract.py  04_run.py  run_all.py  run_rx.py
|       |   `-- tic_common.py
|       `-- config/
|           |-- target_codes.csv          <- the procedure menu (edit to change scope)
|           |-- target_providers.csv       <- optional provider filter
|           `-- payers/bcbstx.json         <- per-payer URLs (add a file per new payer)
|-- 01_raw/
|   |-- npi_tx_xwalk.csv          Texas NPIs + ZIP + region
|   |-- medicare_benchmark.csv    code -> Medicare allowed $
|   `-- _extracted/               prices.csv, providers.csv.gz, rx_prices.csv, inventory.csv
|-- 02_cleaned/                  THE DELIVERABLES (.dta)
|   |-- bcbstx_prices.dta                 main price table
|   |-- bcbstx_rx_prices.dta              prescription-drug prices
|   |-- bcbstx_provider_org_directory.dta organizations by network
|   `-- bcbstx_medicare_benchmark.dta     Medicare amounts
|-- 03_mapping/                  Census shapefiles for the maps:
|                                county, Senate/House districts (+ _shp), ZIP centroids
|-- 03_output/                   figures (.png) + summary tables (.csv)
`-- _documentation/
    |-- guide/                  this PDF
    |-- data_dictionary.csv     every variable, plain English
    `-- website/                the HTML site (index.html)
```

Why Python *and* Stata?

The source files are compressed JSON that must be *streamed*; Stata cannot do that, but Python can. So Python does the download + filter + collapse, leaving small CSVs; Stata then does everything an analyst cares about (clean, label, analyze). `100_data_download.do` is the bridge — it calls the Python pipeline for you. You never have to run Python by hand unless you want to.

To run everything (in Stata, from `_code/`): `do 000_controlfile.do` once, then `100`→`600` in order, or rebuild the website with `do "${docs}/_runall.do"`. The quarterly setup steps (Texas NPI list, region crosswalk, Medicare benchmark) are documented at the top of `100_data_download.do`.

12 Variable reference

The complete, authoritative dictionary — every variable, a plain-English description, and allowed values — is `_documentation/data_dictionary.csv`. The labels applied in `300_labels.do` use identical wording. If you rename a variable, change it in the data, the labels, and the dictionary together.

13 Next steps planning and critical internal questions

This dataset will outlive its first run, and a policy team may ask it to do many things (eg feed a public product, support testimony, brief funders, etc) The answers it might provide differ by question. Below are some possible next steps or input needed to move forward, each tagged **Feasible now**, **Feasible with effort**, **Needs external data**, or **Not feasible (today)**, with a short why.

13.1 Products, policy, and communications

"Can we build a public price-comparison *dashboard*, like our hospital price-transparency dashboard?"
Feasible now (with framing). We already have median prices by service, network, region, and provider type, plus working interactive ; prototypes (the `statashiny` table and the `sparkta` explorers on the analysis page). The constraint is the label: these are prices *published*, not *paid*, and (for now) one insurer. A public product should foreground price *variation* and the facility-vs-Medicare story, and state the ghost-rate caveat plainly.

"Can we say what Texans actually *pay*, total *spending*, or the savings from a proposed policy?"
Not feasible (today). These files carry prices but no *volume* ; how often each service is billed. Spending = price × quantity, and quantity is absent (Section 10.4). Any spending, budget-impact, or savings figure needs an external volume source; without it we can describe price *levels and variation* only.

"Can we put a specific number in legislative testimony?"
Maybe using this data (once verified more carefully and combined with other company TiC datasets now ... Examples of things we could vet and likely speak to: (a) identical shoppable services carry a several-fold price range across Texas (dispersion); (b) facility and inpatient

rates run *above* Medicare (median ≈108%, up to ≈148% for major stays). Treat the *professional* % of Medicare cautiously ; it is dragged below 100% by ghost rates (Section 8) ; and always say "published rates," not "what the plan pays."

"Can we write a blog post to funders around a striking finding?"
Feasible now. The cleanest hooks are price *variation* (the same service, wildly different negotiated prices) and the facility-above / professional-below-Medicare puzzle (Section 8) ; a genuinely novel, defensible finding ; provided it is framed as published rates with the ghost-rate nuance stated.

"Which hospitals or systems are the most expensive ; can we speak to consolidation / market power?"
Needs a re-pull plus external data. Provider *identity* is collapsed off the price rows; naming systems needs a re-extraction that keeps the provider group, plus an NPI→system / hospital-CCN crosswalk. True market-power measures (HHI) additionally need volume-based market shares. Named-provider attribution is feasible with effort; HHI-vs-price is not, today.

"Can we track prices over time ; is transparency 'working'?"
Feasible with effort (set it up now). Each monthly pull is a snapshot keyed on `run_date`; the analysis turns on time-trend panels automatically once two or more are stacked. The investment is simply running the pipeline monthly and appending.

"Can we relate price to *quality* (a 'value' view)?"
Needs external data. Join CMS Hospital Compare / star ratings on a provider crosswalk ; feasible for facilities once the provider-identity re-pull exists, not from these files alone.

"Can we map prices to a specific legislator's district for a member briefing?"
Approximate now; exact needs a re-pull. We produce an *approximate* office-visit price map by Senate and House district and *exact* provider-supply maps by district (Section 9). An exact *price*-by-district map needs a district-keyed re-extraction.

"What is the highest-value/return next step for this work?"
One answer: volume data. A claims/utilization source (the Texas All-Payer Claims Database is the natural fit, tying into Texas 2036's broader APCD work) simultaneously unlocks spending, a *paid*-like % of Medicare, market shares, and ghost-rate down-weighting. Nothing else moves as many questions from "not feasible" to "feasible" at once.

13.2 Expanding the data: more codes, and more payers

Adding more CPT codes or data elements. Feasible; cheap to specify, not free to run. Add rows to `target_codes.csv` (or a field to the extractor) and re-pull ; Section 7.1 has the mechanics. *Pros:* broadens the menu of services you can speak to, at no statistical cost to the codes already in the data (adding codes is *scope*, not *bias*). *Cons:* every addition triggers a full ∼12 GB re-download and rebuild, because the filter runs during extraction; and a bigger menu does not, by itself, make the data more *representative* of all care ; it remains a purposive shoppable-services sample. *Recommendation:* expand in batches tied to a specific policy question, folded into the routine monthly pull rather than done ad hoc.

Running other Texas insurers (UnitedHealthcare, Cigna, Anthem-type plans) and combining them. Feasible with effort, high value ; with real limits. This is the most consequential expansion, so in detail:

1. **Pull each payer *separately.*** Each insurer publishes its own Transparency-in-Coverage index. Adding one is a new `config/payers/<payer>.json` (its index and file-list endpoints) run through the *same* pipeline, producing its own `.dta` files in the same schema. Isolate, validate, and clean each payer first ; do not pull them together.
2. **Then stack the comparable subset.** Because every payer uses the CMS TiC format, the columns align. `append` the per-payer price files into one panel with a new `payer` column. Compare on the *common, comparable* rows only: same `procedure_code`, same `provider_type`, `dollar_ffs`, and ideally the same region (better, the same providers). Compare medians by code/region across payers ; never a raw pooled average, which would just reflect which payer published more rows.
3. **Why it is worth it.** Three major commercial payers cover a large share of the Texas commercial market, so stacking them yields a far more *representative* statewide price level *and* the policy story this data is uniquely good at: the *same service, same provider, different payer, different negotiated price*.
4. **What is still limited.** Stacking adds no volume, so spending and market shares stay out of reach; ghost rates afflict every payer's file (at differing rates; though we could empircally study these rates adn the sensitivity of the cutoff value we selected (25%)); network and coverage-segment definitions differ across payers, so "comparable" must be enforced on code + provider, not on plan; and it still excludes Medicaid, Medicare Advantage, and small payers. It is a better *commercial-price* picture of Texas ; not a complete one.