



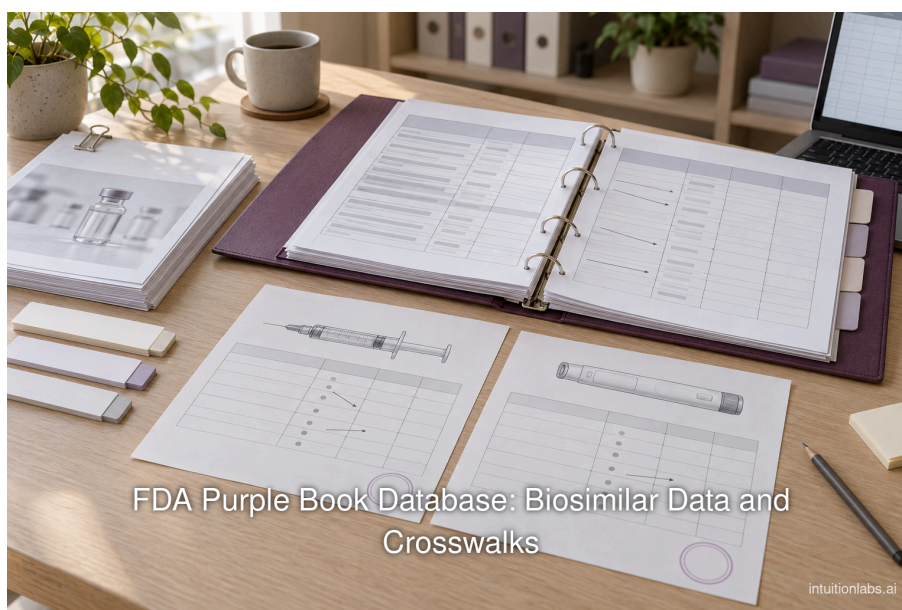
INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

FDA Purple Book Database: Biosimilar Data and Crosswalks

Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-10-15 · fda purple book database · fda purple book · biosimilars · interchangeable biosimilars · reference products · biologics data · purple book csv · regulatory intelligence · biologics crosswalk

Original article: <https://intuitionlabs.ai/articles/fda-purple-book-data-guide>



In this article

1. Executive Summary
 2. Introduction and Background
 3. Research Methodology and Source Boundaries
 4. Entities, Relationships and Field Dictionary
 5. Verified Reference-Product Crosswalk
 6. Extraction, Stable Keys and Change Detection
 7. Joins and Decision Coverage
 8. Data Analysis and Evidence
 9. Implications and Future Directions
 10. Conclusion
-

Executive Summary

The **FDA Purple Book database**, maintained by the United States Food and Drug Administration (FDA), is the starting point for identifying licensed biological products, their reference-product relationships, and biosimilar or interchangeable evaluations. It covers Center for Drug Evaluation and Research (CDER) biologics and Center for Biologics Evaluation and Research (CBER) allergenic, cellular and gene therapy, hematologic, and vaccine products. Its scope is therefore broader than a biosimilar approval list. (purplebooksearch.fda.gov) Federal law limits a biosimilar application to one reference product, making a directed relationship model more appropriate than a flat list of equivalent brands. (www.govinfo.gov)

Original analysis of the **September 2026** full monthly snapshot found **2,251 product rows**, **864 distinct biologics license applications (BLAs)**, and no duplicate BLA/product-number pairs.

(www.accessdata.fda.gov) The records included **1,197 CDER** and **1,054 CBER** rows, with **146 interchangeable** and **100 biosimilar** section 351(k) product rows under the Public Health Service Act.

(www.accessdata.fda.gov) These presentation-level counts are distinct from application counts and from the **89 displayed rows** counted in FDA's separate biosimilar table during verification on **October 15, 2026**.

(www.fda.gov) Every published count needs its unit, source period, and status filter. A composite key can validate row uniqueness within a snapshot without establishing permanent identity across every future release. (specs.frictionlessdata.io)

The worked crosswalk preserves proper and proprietary names, applicant, application/product identifiers, reference names, evaluation, dates, exclusivity fields, source links, and verification dates. **Wezlana** illustrates why the product grain matters: its selected syringe and autoinjector presentations carry different evaluations. (purplebooksearch.fda.gov) Reference application identifiers are curated joins supported by presentation evidence, rather than supplied numeric reference keys. The resulting mapping should retain its source and derivation history. (www.w3.org) Approval and interchangeability do not independently answer operational substitution or reimbursement questions: state pharmacy requirements and quarter-specific Centers for Medicare & Medicaid Services (CMS) payment information remain separate evidence layers. (www.fda.gov) (www.cms.gov)

For extraction, isolate the monthly report's full-snapshot section from its changes section before counting; otherwise changed products can be counted twice. (purplebooksearch.fda.gov) Preserve raw strings, distinguish missing dates from negative determinations, version reference-resolution decisions, and reconcile accepted snapshots. Declared null handling is essential before type conversion. (specs.frictionlessdata.io) Join labels by their own identifiers and versions, with DailyMed publication dates kept separate from regulatory approval dates. (dailymed.nlm.nih.gov) The recommended output is an auditable relationship dataset whose labeling, package, payment, launch, patent, and substitution conclusions each retain the additional evidence needed for that particular decision.

Introduction and Background

The **FDA Purple Book database** is a regulatory relationship source for biological products. Its coverage includes biological products regulated by the **Center for Drug Evaluation and Research (CDER)** and allergenic, cellular and gene therapy, hematologic, and vaccine products regulated by the **Center for Biologics Evaluation and Research (CBER)**. A search for biosimilars therefore addresses only part of the database. (purplebooksearch.fda.gov)

For portfolio teams, the useful question is usually more specific than whether a brand appears: which application and presentation does a record describe, what reference product supports its relationship, and what determination applies to that presentation? Federal law specifies **one reference product per biosimilar application**. A data model should preserve that directed relationship instead of treating shared ingredient names as proof that every product is interchangeable with every other product. (www.govinfo.gov)

The database also has a practical audience beyond regulatory affairs. The Centers for Medicare & Medicaid Services (**CMS**) includes the Purple Book among pharmacist resources. Nevertheless, reimbursement analysis needs additional sources because CMS publishes Medicare Part B payment limits separately, by quarter. (www.cms.gov) (www.cms.gov)

This report is written for regulatory intelligence, biologics strategy, market access, and data engineering teams. IntuitionLabs lists data pipelines, integration, warehousing, and business intelligence among its services; the implementation perspective here is to preserve source meaning before using regulatory records in analytics. (intuitionlabs.ai) The proposed schema, reconciliation rules, and controls below are analyst recommendations, rather than FDA database specifications. Their provenance approach follows the World Wide Web Consortium (**W3C**) distinction between source entities and the activities that produce derived data. (www.w3.org)

The publication reference date is **October 13, 2026**. Research verification occurred on **October 15, 2026**. Quantitative results use the **September 2026 monthly snapshot**, while record checks are explicitly dated when performed. These are observation dates, not assertions that a later page was observed earlier. Retaining separate source, retrieval, and transformation timestamps makes that distinction reproducible under a provenance model. (www.w3.org)

Research Methodology and Source Boundaries

Start with the decision

Use the following routing rules before collecting a portfolio dataset:

- **FDA relationship:** Identify the biosimilar application's reference product, preserving the direction of the statutory relationship. (www.govinfo.gov)
- **Substitution:** Obtain the FDA determination and consult current local pharmacy requirements before operational use. FDA describes pharmacy substitution as subject to state pharmacy laws. (www.fda.gov)
- **Labels:** Retrieve the specific product's labeling and version history instead of copying label content across a reference relationship. (dailymed.nlm.nih.gov)
- **Payment:** Select the relevant CMS payment quarter and billing code before adding a reimbursement measure. (www.cms.gov)
- **Packages:** Resolve National Drug Codes (**NDCs**) and package units independently of an application-level relationship. The Social Security Act explicitly associates units and package size with NDCs. (www.ssa.gov)
- **Auditability:** Preserve the source artifact, transformation activity, and resulting dataset as distinct provenance objects. (www.w3.org)

The source hierarchy used here places FDA records and downloads first for US product facts, federal legislation first for statutory definitions, and CMS or National Library of Medicine (**NLM**) documentation first for their respective enrichment datasets. This is an evidence allocation rule: each source answers the question it actually governs. W3C's provenance model provides a useful vocabulary for documenting how those sources contribute to a derived result. (www.w3.org)

Keep jurisdiction attached to every determination

International sources help expose a common modeling error: translating a familiar term into a US status without its jurisdiction. The **European Medicines Agency (EMA)** describes EU-approved biosimilars as interchangeable from a scientific viewpoint, while pharmacy substitution decisions remain with Member States. Those statements should not populate an FDA interchangeability field. (www.ema.europa.eu)

The **World Health Organization (WHO)** offers biosimilar evaluation principles that national regulators can adopt. Its guidelines exclude vaccines and plasma-derived products, so their scope cannot define the coverage of a register containing a broader set of biological products. (www.who.int) (www.who.int) Health Canada likewise describes a reference biologic relationship and specifies that a biosimilar cannot itself serve as the reference for a subsequent biosimilar. (www.canada.ca)

Australia's **Therapeutic Goods Administration (TGA)** requires an Australian reference medicine and describes bridging evidence when certain studies use a foreign comparator. Record the jurisdiction and role of a comparator rather than interpreting a foreign study comparator as the US reference identity. (www.tga.gov.au) (www.tga.gov.au) These boundaries keep a global portfolio useful without importing foreign authorization decisions into the US relationship graph.

Entities, Relationships and Field Dictionary

Separate applications, presentations and reference assertions

An application, a product presentation, a license holder, and a reference-product assertion should be separate entities. FDA's user guide explicitly states that a single biosimilar application can contain both biosimilar and interchangeable products. It also describes core-name grouping in Simple Search and fuller information in Advanced Search, with selected-column exports to comma-separated values (CSV), Microsoft Excel, or Portable Document Format (PDF). Consequently, a search-results group is a discovery aid, rather than a sufficient database key. (purplebooksearch.fda.gov)

The recommended relationship graph is below. A reference assertion belongs to an application at an observation time and points to a curated reference entity. Application identifiers used to document that reference entity are retained as evidence. This preserves the statutory single-reference rule without assuming a displayed reference brand is a unique application identifier. (www.govinfo.gov) Foreign keys can then enforce the relationships in the internal database; Table Schema documents the same general relational concept. (specs.frictionlessdata.io)

```
erDiagram
    APPLICATION ||--o{ PRODUCT : contains
    PRODUCT ||--o{ OBSERVATION : has
    SOURCE_SNAPSHOT ||--o{ OBSERVATION : supplies
    APPLICATION ||--o{ REFERENCE_ASSERTION : has_at_observation_time
    REFERENCE_PRODUCT ||--o{ REFERENCE_ASSERTION : is_target
    REFERENCE_PRODUCT ||--o{ REFERENCE_RECORD : documented_by
    SOURCE_SNAPSHOT ||--o{ REFERENCE_RECORD : supports
```

Table 1 groups the monthly download's field labels into an implementable dictionary. The interactive table separately exposes the corresponding approval, interchangeability, and reference-name fields. Preserve original column labels in the landing layer even if warehouse names differ. (purplebooksearch.fda.gov)

Field group	Source labels	Recommended interpretation and storage
Identity	BLA Number; Product Number; License Number; Applicant	Keep identifiers as strings. Distinguish application, product, and applicant/license information.
Naming	Proprietary Name; Proper Name	Retain exact source spelling alongside normalized search values. Neither name alone is a primary key.
Product description	Strength; Dosage Form; Route of Administration; Product Presentation; Center	Preserve presentation and route so relationship resolution can consider the actual product.
Evaluation	License Type	Preserve the product's source category, including the biosimilar or interchangeable distinction.

Field group	Source labels	Recommended interpretation and storage
Record status	Marketing Status; Licensure	Store separately. Historical inclusion and source marketing categories answer different questions.
Relationship	Ref. Product Proper Name; Ref. Product Proprietary Name	These are supplied names. A resolved reference application identifier is an additional, documented join.
Dates and submission	Approval Date; Inter. Approval Date; Supplement Number; Submission Type; Inter. Supplement Number	Preserve approval and interchangeability dates independently, with submission context.
Exclusivity	Date of First Licensure; Exclusivity Expiration Date; First Interchangeable Exclusivity Exp. Date; Ref. Product Exclusivity Exp. Date; Orphan Exclusivity Exp. Date	Keep each field separate and retain non-date source values.
Change and patent metadata	N/R/U; Patent List Provided	Separate change markers from snapshot attributes. A supplied patent-list flag is a source field, not a freedom-to-operate conclusion.

The dictionary is intentionally descriptive. It does not claim that a field name guarantees a complete determination. FDA states that it has not determined first licensure for every relevant biological product; its FAQ also distinguishes an undetermined exclusivity period from a negative eligibility conclusion and explains that transition products retain their initial approval dates. Therefore, a missing date should remain an unresolved source state. (purplebooksearch.fda.gov)

In a typed warehouse, use a date value plus a raw-value and interpretation-status field. Frictionless Table Schema converts configured missing-value strings before type conversion, illustrating why an importer must decide explicitly how blanks are treated. (specs.frictionlessdata.io) Its discussion also recognizes that missing values have different origins. Preserve distinctions such as blank, an explicit source determination, and an unparseable value rather than coercing all three into the same business conclusion. (specs.frictionlessdata.io)

Biosimilar versus interchangeable

Interchangeability has a specific US statutory meaning involving substitution for the reference product without intervention by the prescribing healthcare provider. That definition does not establish unrestricted equivalence among every member of a reference group. (www.govinfo.gov) EMA's separate scientific description can include using another biosimilar of the same reference product, demonstrating why a globally shared Boolean named interchangeable loses necessary context. (www.ema.europa.eu)

Reference-product exclusivity is another separate concept. The statute generally establishes a **12-year** period tied to first licensure before a biosimilar approval can become effective, subject to its qualifications. Do not generate a definitive expiry merely by adding that interval to whichever approval date happens to be populated. (www.govinfo.gov) For legal decisions about a particular program, regulatory counsel should examine the applicable determination and statutory provisions.

Verified Reference-Product Crosswalk

Table 2 presents a selected crosswalk rather than a complete FDA Purple Book biosimilar list. Each row uses the September snapshot and was checked against retrieved official record or Advanced Search content on **2026-10-15**. Application and product identifiers are shown separately; blank means the supplied field is empty. **FI** below means the **First Interchangeable Exclusivity Expiration Date** field, not reference-product exclusivity.

Proper / proprietary name	Applicant; license number	BLA / product	Presentation; product type	Reference product; resolved BLA	Biosimilar / interchangeable evaluation	Approval; interchangeability approval	Exclusivity field present	Official record/source; verified on
adalimumab / Humira	AbbVie Inc.; 1889	125 057 / 001	40MG/0.8 ML autoinjector; 351(a)	No reference name supplied on this row	351(a), reference anchor for the selected adalimumab relationship	2002-12-31; blank	Orphan field: 2028-02-24; reference-product field blank	(purplebo oksearch. fda.gov) 2026-10-15
adalimumab-atto / Amjevita	Amgen Inc.; 1080	761 024 / 001	20MG/0.4 ML prefilled syringe; 351(k)	Humira / 125057	Interchangeable	2016-09-23; 2024-08-20	FI blank	(purplebo oksearch. fda.gov) 2026-10-15
ustekinumab-auub / Wezlana	Amgen Inc.; 1080	761 285 / 001	45MG/0.5 ML prefilled syringe; 351(k)	Stelara / 125261, subcutaneous reference mapping	Interchangeable	2023-10-31; 2023-10-31	FI: 2025-04-30	(purplebo oksearch. fda.gov) 2026-10-15
ustekinumab-auub / Wezlana	Amgen Inc.; 1080	761 285 / 004	45MG/0.5 ML autoinjector; 351(k)	Stelara / 125261, subcutaneous reference mapping	Biosimilar	2024-12-26; blank	FI blank	(purplebo oksearch. fda.gov) 2026-10-15
insulin glargine / Lantus	Sanofi-Aventis U.S. LLC; 1752	021 081 / 001	1,000UNIT S/10ML multidose vial; 351(a)	No reference name supplied on this row	351(a), reference anchor for the selected insulin relationship	2000-04-20; blank	FI and reference-product fields blank	(purplebo oksearch. fda.gov) 2026-10-15
insulin glargine-yfgn / Semglee	Biocon Biologics Inc.; 2324	761 201 / 001	1,000UNIT S/10ML multidose vial; 351(k)	Lantus / 021081	Interchangeable	2021-07-28; 2021-07-28	FI: 2022-11-15	(purplebo oksearch. fda.gov) 2026-10-15

The Wezlana rows demonstrate the practical consequence of product-level evaluation: copying the syringe's designation onto the autoinjector would change the source meaning. The selected subcutaneous Stelara mapping is documented against its official reference record. It is an analyst-resolved identifier, since the

supplied reference columns contain names rather than a numeric foreign key. (purplebooksearch.fda.gov)

Record supplied names, resolved identities, resolution methods, and evidence in the crosswalk. (www.w3.org)

Apply the relationship-validation controls described above. (specs.frictionlessdata.io)

Names should therefore support candidate matching, followed by application and presentation checks.

RxNorm's normalized drug names combine ingredient, strength, and dose form, which is useful for a different

matching task from establishing a regulatory reference relationship. (www.nlm.nih.gov) TGA's use of an

Australian biological name without a specific suffix also shows why a US suffix rule should not be exported into a universal biologic identifier. (www.tga.gov.au)

Publish source fields, resolved identifiers, unresolved reasons, and retrieval dates together; distinguish absent

source values from unfinished joins. (specs.frictionlessdata.io) Retain earlier mapping releases for

reproducibility. (www.w3.org) (www.w3.org)

Extraction, Stable Keys and Change Detection

Download the source before normalizing it

For **FDA Purple Book data download** workflows, distinguish the monthly CSV or Excel report from a selected Advanced Search export. The monthly report starts with changes and then supplies the full snapshot. Its change codes distinguish updates, newly approved products, and previously approved products added in the release. Concatenating both sections creates duplicate observations of changed rows.

(purplebooksearch.fda.gov)

The listed CSV artifact could not be read during this session after both permitted retrieval paths were attempted. The Excel equivalent was accessible and supplies the calculations here. This is a retrieval limitation of this analysis, rather than evidence that FDA does not provide CSV data. No documented public bulk application programming interface (API) was identified in the official materials reviewed; the reproducible extraction routes examined here are monthly files and interface exports.

The ingestion sequence should be explicit:

- **Archive:** Retain the source URL, filename, retrieval timestamp, and a content hash. Treat the source artifact as a provenance entity. (www.w3.org)
- **Identify sections:** Locate the full-snapshot header structurally. Do not assume a fixed line number will remain valid in future releases. Header names provide the field correspondence. (www.rfc-editor.org)
- **Parse correctly:** Use a CSV parser that handles quoted commas, line breaks, and embedded double quotes. (www.rfc-editor.org) (www.rfc-editor.org)
- **Validate shape:** Compare the parsed field count with the observed header and quarantine structurally inconsistent records. (www.rfc-editor.org)
- **Preserve keys:** Keep application, product, and license identifiers as strings, with a documented composite-key rule. Table Schema permits a primary key comprising multiple fields. (specs.frictionlessdata.io)
- **Preserve source values:** Keep the raw exclusivity cell alongside its typed representation and interpretation status. Missing-value conversion belongs in an explicit schema. (specs.frictionlessdata.io)

- **Record derivations:** Associate the parser version and mapping release with the dataset they produced. (www.w3.org)
- **Reconcile:** Compare a rebuilt full snapshot with the prior accepted snapshot before releasing portfolio metrics. Store the comparison as another derived artifact. (www.w3.org)

For the supplied download, the candidate operational key is **BLA Number plus Product Number**, scoped to a source observation. This is a tested property of the examined snapshot, not a claim that FDA guarantees permanent identifiers. A composite primary key can enforce uniqueness within an accepted dataset, while an observation key distinguishes later versions. (specs.frictionlessdata.io) Keep rejected rows and their reasons outside the accepted table so quality failures remain inspectable.

An appropriate representation separates identity from observation. For example, an internal product identifier may point to application and product-number strings, while an observation carries source license type, dates, applicant, and status. W3C defines an entity through fixed aspects, providing a rationale for preserving an earlier observation rather than silently overwriting it. (www.w3.org) A transformed warehouse row can explicitly derive from that observation. (www.w3.org)

Compare meanings, not just files

A source file hash detects any byte change; a normalized row comparison identifies the affected fields. These controls answer different engineering questions. A proposed change record should identify the old and new source observations, the changed columns, the ingestion activity, and whether a downstream mapping was recomputed. That is an application of provenance rather than a new FDA status classification. (www.w3.org)

Prioritize changes to license type, reference names, applicant, presentation, dates, and the interpretation of exclusivity cells. Preserve unknown source fields when a schema evolves, then decide deliberately whether they belong in the accepted model. CSV header correspondence and consistent record width support that structural check. (www.rfc-editor.org) (www.rfc-editor.org) Do not let a new column shift positional mappings unnoticed.

For crosswalk revisions, distinguish a changed source fact from a changed matching decision. A reference name may remain identical while an analyst corrects the resolved application. Record the second event in the mapping release and retain the original evidence. Foreign-key validity and provenance are complementary controls: one checks relational structure, while the other explains how a value was produced.

(specs.frictionlessdata.io) (www.w3.org)

Joins and Decision Coverage

Enrich at the correct grain

Drugs@FDA and **openFDA** can enrich application information, but their identifiers should be resolved explicitly. The openFDA Drugs@FDA documentation describes additional harmonized identifiers as annotations available for an application number when present. A separate documentation page cautions that not all records have harmonized fields. A failed harmonized join is therefore a matching outcome to investigate, rather than evidence that the biological product lacks approval. (open.fda.gov) (open.fda.gov)

For labeling, preserve the **Structured Product Labeling (SPL)** set identifier, version, and publication date separately. DailyMed exposes label version history by SET ID and NDCs associated with that identifier. (dailymed.nlm.nih.gov) (dailymed.nlm.nih.gov) Its published_date is the date of publication on DailyMed. That field should not overwrite an FDA approval date or become the effective date of a reference relationship. (dailymed.nlm.nih.gov)

RxNorm adds a terminology layer. Its normalized names describe ingredient, strength, and dose form, while its documentation says RxNorm concept unique identifiers (**RXCUIs**) are not reused. Retain the terminology release and mapping evidence alongside the code so a stable identifier does not erase the history of a matching decision. (www.nlm.nih.gov) (www.nlm.nih.gov)

External joins need separate validation rules:

- **Application normalization:** Medicaid's explicitly dated August 2018 dictionary describes a seven-digit FDA application field. Preserve its raw representation and document conversion to another source's format. (www.medicaid.gov)
- **Labeler identity:** That dictionary defines the NDC labeler segment. Treat it as its own identifier rather than assuming it is identical to the Purple Book applicant field. (www.medicaid.gov)
- **Package units:** Resolve the NDC and its package size before aggregating package observations. (www.ssa.gov)
- **Billing units:** Keep the payment-code quantity separate from package and presentation counts. (www.ssa.gov)
- **Indications:** Validate against the selected product's label and retain the retrieved version, rather than inheriting label content through a reference join. (dailymed.nlm.nih.gov)
- **Label chronology:** Preserve DailyMed publication time as a labeling observation timestamp. (dailymed.nlm.nih.gov)
- **Payment period:** Attach the applicable quarter to a reimbursement observation. (www.cms.gov)
- **Terminology release:** Retain the release and evidence used for each RxNorm identifier mapping. (www.nlm.nih.gov)

The Medicaid dictionary is a historical specification, not a verified current schema contract. Its Market Date warning is still a useful semantic caution: a drug's market date may precede marketing by a particular labeler. Do not promote it into a current license holder's launch date without additional evidence. (www.medicaid.gov)

Which questions can the database answer?

Table 3 separates regulatory relationship questions from operational decisions. The supplemental sources identify the next evidence layer; none turns a Purple Book row into a universal commercial conclusion.

Question	Purple Book contribution	Additional evidence or control
FDA status and reference relationship	Source license type, reference names, application/product identifiers, and dates	Resolve the directed relationship rather than equating all group members. (www.govinfo.gov)
Pharmacy substitution	FDA evaluation is relevant input	Current state requirements and workflow conditions; consult pharmacy/regulatory specialists. (www.fda.gov)

Question	Purple Book contribution	Additional evidence or control
Commercial launch or availability	A regulatory record and source marketing-status value	Independent dated commercial evidence; even a separate Market Date has source-specific semantics. (www.medicaid.gov)
Patents and freedom to operate	Supplied patent-list and exclusivity fields, where populated	Product-specific legal analysis. Reference-product exclusivity is a specific statutory mechanism. (www.govinfo.gov)
Reimbursement	Identity and relationship inputs	Program, quarter, billing code, eligibility, and payment units. (www.cms.gov) (www.ssa.gov)
Product-specific labeling	A starting identity for retrieval	Retrieve the product's label and version history. (dailymed.nlm.nih.gov)
Foreign interchangeability	US record only	Jurisdiction-specific evaluation and substitution rules. (www.ema.europa.eu) (www.ema.europa.eu)

In portfolio dashboards, leave launch, coverage, and legal-clearance cells unresolved until supporting evidence is available. (specs.frictionlessdata.io) Attribute accepted enrichments to their sources. (www.w3.org)

For payment analytics, **average sales price (ASP)** is a separate measure with its own policy context. CMS describes a temporary methodology for qualifying biosimilars of their ASP plus **8% of the reference biological product's ASP**. Qualification and the applicable period must be established before applying that formula. (www.cms.gov) A reusable model should therefore keep the billing unit and payment quarter outside the FDA relationship entity. (www.ssa.gov) (www.cms.gov)

Data Analysis and Evidence

Snapshot counts and denominators

Original analysis of the **September 2026** monthly Excel full-snapshot section produced **2,251 product rows**, **864 distinct BLA numbers**, and **2,251 distinct BLA/product-number pairs**. (www.accessdata.fda.gov) The center breakdown was **1,197 CDER** and **1,054 CBER** rows. (www.accessdata.fda.gov) License types comprised **2,005 section 351(a)** rows, **146 interchangeable section 351(k)** rows, and **100 biosimilar section 351(k)** rows. (www.accessdata.fda.gov) These are calculated counts of the downloaded records, rather than FDA-published headline approval totals.

The calculation isolated records after the second field-header occurrence, retained rows matching the supplied field structure, and counted application and application/product identifiers as strings. Only the full-snapshot section was counted. The snapshot also contains historical licensure categories, including voluntarily revoked and revoked records, so an all-row total is not a count restricted to currently licensed products. Categories are reported here only as aggregate database semantics, without attributing events to named products.

The **246 section 351(k) product rows** span **95 distinct BLAs**, **21 reference proper names**, and **23 reference proprietary names** in that same snapshot. (www.accessdata.fda.gov) These different denominators explain why presentation, application, and reference-family counts must be labeled. Keep reference-name counts separate from resolved application counts, using the resolution controls described above. (specs.frictionlessdata.io)

For reproducibility, the calculation can be expressed independently of a particular spreadsheet library:

```
# full_snapshot is the isolated lower section, with identifier strings preserved.
rows = list(full_snapshot)
product_keys = {(r["BLA Number"], r["Product Number"]) for r in rows}
application_keys = {r["BLA Number"] for r in rows}
assert len(product_keys) == len(rows)
k_rows = [r for r in rows if r["License Type"].startswith("351(k)")]
counts_by_type = Counter(r["License Type"] for r in rows)
```

This is calculation pseudocode, not an extraction API. A production parser must identify headers, preserve quoted fields, and validate record width before running it. Request for Comments (RFC) 4180 documents those CSV structural concerns; it is an informational description, not evidence of an FDA API commitment. (www.rfc-editor.org) (www.rfc-editor.org)

Why the biosimilar table count differs

The separately fetched FDA Biosimilar Product Information table displayed **89 rows** on the verification date. (www.fda.gov) This is an analyst count of displayed rows, including entries that combine proprietary names. It is not the number of all licensed biologics, all Purple Book applications, or all biosimilar presentations.

The comparison does not establish missing approvals or explain every difference between publication surfaces. Its sources use different grains and observation periods. A reconciled dataset should document the unit counted, status filter, snapshot date, and handling of combined names before presenting any discrepancy as a portfolio finding. Retain the reconciliation inputs and output. (www.w3.org) (www.w3.org)

WHO's guideline scope supplies another denominator warning: it excludes vaccines and plasma-derived products, whereas an all-biologics register can be broader. (www.who.int) A measure titled simply "biologics" is insufficiently specified unless its regulatory scope and unit of analysis accompany the result.

Implications and Future Directions

The practical objective is a maintained, evidence-bearing relationship model. Regulatory intelligence can identify candidate reference groups; strategy teams can count competing applications or presentations at a declared grain; data engineering can reproduce mappings; and market access can attach program-specific reimbursement information. Payment enrichment still needs its own quarterly source and billing-unit definition. (www.cms.gov) (www.ssa.gov)

For an advisory implementation perspective, IntuitionLabs describes connecting AI to authoritative enterprise sources with retrieval, citations, and evaluation. Those same controls are useful when regulatory data feed analytical or AI workflows. (intuitionlabs.ai) Keep source records, accepted joins, and generated outputs separately attributable. (www.w3.org)

A monthly refresh, supplemented by checks following relevant new approvals, should use this release checklist:

- **Freeze the input:** Retain the source artifact and retrieval metadata before transformation. (www.w3.org)

- **Check schema drift:** Compare field names and record width with the accepted parser contract. (www.rfc-editor.org)
- **Recheck identity:** Confirm candidate-key uniqueness and investigate duplicate application/product pairs. (specs.frictionlessdata.io)
- **Review determinations:** Compare product-level evaluations and dates with the prior accepted observation.
- **Review null semantics:** Preserve unresolved or non-date source states through type conversion. (specs.frictionlessdata.io)
- **Reconcile labels:** Check the relevant SET ID version history instead of assuming an unchanged label. (dailymed.nlm.nih.gov)
- **Refresh payment separately:** Select the applicable quarter and retain billing-unit context. (www.cms.gov) (www.ssa.gov)
- **Publish evidence:** Version the crosswalk, resolution decisions, and transformation provenance together. (www.w3.org)

Global portfolios should retain jurisdiction as a first-class field. WHO principles support national frameworks, while Health Canada authorization requires a submission and TGA reference requirements relate to Australian registration. Those authorities provide context and separate evidence layers, not replacement FDA determinations. (www.who.int) (www.canada.ca) (www.tga.gov.au) Canadian patent and data-protection requirements likewise belong in their own jurisdictional analysis. (www.canada.ca)

Make unresolved reference keys, unverified launch dates, and unavailable coverage evidence visible. Apply the key-validation controls above without treating a valid join as a complete interpretation. (specs.frictionlessdata.io) (specs.frictionlessdata.io)

Conclusion

The FDA Purple Book database is most useful when its product and relationship semantics survive extraction. The implementation should preserve application and presentation identity, directed reference assertions, source dates, and separate evaluation fields. A legally defined reference relationship has a narrower meaning than an unrestricted equivalence group. (www.govinfo.gov) The statutory definition of interchangeability concerns substitution for the reference product without prescriber intervention. (www.govinfo.gov)

The verified crosswalk demonstrates why presentation-level evaluation matters, while the snapshot analysis demonstrates why product-row totals, application totals, and a separately published biosimilar table are different measures. The recommended database retains those denominators and observation dates instead of collapsing them into one count. Retain the inputs and transformations behind each result. (www.w3.org)

Downstream enrichment should remain explicit. Label histories, terminology mappings, NDC packages, payment units, and reimbursement periods belong to their respective data layers. DailyMed's version-history interface and the statutory billing-unit definition illustrate the different grains involved. (dailymed.nlm.nih.gov) (www.ssa.gov) A relationship crosswalk becomes more useful when each attached measure carries its own source and meaning.

For recurring use, publish accepted mappings with reproducible inputs, unresolved cases, and a clear refresh process, using the extraction controls above. (specs.frictionlessdata.io) (specs.frictionlessdata.io) The result is a regulatory data asset that supports portfolio questions while keeping launch, substitution, patent, and reimbursement conclusions tied to the additional evidence each requires.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. <https://purplebooksearch.fda.gov/about>
2. <https://www.govinfo.gov/link/uscode/42/262>
3. https://www.accessdata.fda.gov/drugsatfda_docs/PurpleBook/2026/purplebook-search-September-data-download.xlsx
4. <https://www.fda.gov/drugs/biosimilars/biosimilar-product-information>
5. <https://specs.frictionlessdata.io/table-schema/>
6. <https://purplebooksearch.fda.gov/index.cfm?blaNo=761285&event=productdetails>
7. <https://www.w3.org/TR/prov-dm/>
8. <https://www.fda.gov/news-events/press-announcements/fda-approves-interchangeable-biosimilar-multiple-inflammatory-diseases>
9. <https://www.cms.gov/medicare/payment/fee-for-service-providers/part-b-drugs/average-drug-sales-price>
10. <https://purplebooksearch.fda.gov/index.cfm?event=downloads>
11. https://dailymed.nlm.nih.gov/dailymed/webservices-help/v2/spls_api.cfm
12. <https://www.cms.gov/pharmacist-information-center>
13. <https://intuitionlabs.ai/>
14. <https://dailymed.nlm.nih.gov/dailymed/app-support-web-services.cfm>
15. https://www.ssa.gov/OP_Home/ssact/title18/1847A.htm
16. <https://www.ema.europa.eu/en/human-regulatory-overview/biosimilar-medicines-overview>
17. <https://www.who.int/publications/m/item/guidelines-on-evaluation-of-biosimilars>
18. <https://www.canada.ca/en/health-canada/services/drugs-health-products/biologics-radiopharmaceuticals-genetic-therapies/biosimilar-biologic-drugs/handbook-healthcare-professionals/about.html>
19. <https://www.tga.gov.au/resources/guidance/registering-biosimilar-medicine-australian-register-therapeutic-goods-artg>
20. <https://purplebooksearch.fda.gov/userguide>
21. <https://purplebooksearch.fda.gov/advanced-search>
22. <https://purplebooksearch.fda.gov/index.cfm?event=faqs>
23. <https://purplebooksearch.fda.gov/index.cfm?blaNo=125057&event=productdetails>
24. <https://purplebooksearch.fda.gov/index.cfm?blaNo=761024&event=productdetails>
25. <https://purplebooksearch.fda.gov/index.cfm?blaNo=021081&event=productdetails>
26. <https://www.nlm.nih.gov/research/umls/rxnorm/overview.html>
27. <https://www.rfc-editor.org/rfc/rfc4180.html>
28. <https://open.fda.gov/apis/drug/drugsfda/understanding-the-api-results/>

29. <https://open.fda.gov/apis/openfda-fields/>
30. <https://www.medicaid.gov/medicaid-chip-program-information/by-topics/prescription-drugs/downloads/recordspecficationanddefinitions.pdf>
31. <https://www.cms.gov/files/document/biosimilar-faqs.pdf>

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore [AI enablement](#), [pharma and biotech software development](#), [data engineering and BI](#), and [Veeva services](#). [Contact IntuitionLabs](#) to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: <https://intuitionlabs.ai>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.