

Supplementary Materials

A Reproducible Common Column Model and Crosswalk for Harmonizing Multi-Country Hospital Discharge Data

Document map

1. Scope and resource overview
2. Repository and artifact structure
3. Canonical column model and field availability
4. Source-to-canonical field mapping and authoritative variable provenance
5. Crosswalk execution model
6. Source-specific preprocessing notes
7. Value normalization rules
8. Technical validation protocol
9. Known anomalies and information-loss map
10. Reproducibility and extension guidance

1. Scope and resource overview

This supplementary document describes the technical implementation of the harmonization resource. It keeps the main article concise while retaining enough operational detail for audit, reproducibility, and extension. The resource centres on five source-specific declarative crosswalks and a shared execution pipeline that maps heterogeneous hospital discharge files to a lightweight canonical model.

The documentation distinguishes three levels of harmonization: file-level handling, column-level mapping, and value-level normalization. It also separates technical validation from epidemiological validation. The purpose of the validation described here is to verify that the pipeline produces files consistent with the declared model and that information loss is transparent, not to establish clinical comparability of outcomes across health systems.

Table S1. Source overview.

Code	Jurisdiction	Source system	Years/files	Input -> output columns	Native encoding/separator	Diagnosis representation	Cost representation
kr	Republic of Korea	NHIS discharge/claims data	2002-2023 (22 files)	17-19 -> 14	EUC-KR (,)	ICD-10	KRW; USD derived
br	Brazil	DATASUS / SIHSUS (AIH)	2006-2025 (20 files)	113 -> 14	UTF-8 (,)	ICD-10	BRL and native USD field
mx	Mexico	SAEH / SSA Egresos Hospitalarios	2010-2025 (16 files)	39-50 -> 10	UTF-8 + BOM (, and)	ICD-10	Not available
cl	Chile	DEIS-MINSAL Egresos Hospitalarios	2001-2025 (25 files)	15-18 -> 7	Latin-1 (,)	ICD-10	Not available

ny	USA, New York State	SPARCS Hospital Inpatient Discharges, de-identified	2009–2024 (16 files)	33–34 → 10	Latin-1 (,)	AHRQ CCS/CCSR proxy	USD; Total Charges
----	---------------------	---	----------------------	------------	-------------	---------------------	--------------------

2. Repository and artifact structure

The release package contains the code, crosswalks, helper assets, validation scripts, validation outputs and source-access records listed in Table S3, together with a machine-readable manifest (RELEASE.json) recording the commit, the tag, the crosswalk version of each source, the pinned and installed dependency versions, the command sequence, and the size and SHA-256 checksum of every released artefact taken individually rather than as a directory digest, covering 54 code artefacts and the deposited data files. Running the manifest in check mode recomputes all checksums and fails if an artefact changed without its version being raised. Where source licences or provider terms prohibit redistribution, raw microdata are replaced by access recipes that allow users to obtain the same public source files from the original providers; the manifest lists what is not included and why.

Table S2. Raw data source access records.

Code	Jurisdiction	Year(s) / files	Raw data source URL	Access date
kr	Republic of Korea	2002–2023 (22 files)	https://www.data.go.kr/data/15007122/fileData.do	[2026-05-29]
br	Brazil	2006–2025 (20 files)	Python library - <code>pysus.online_data.SIH (/ingest/data_downloader.py)</code>	[2026-04-10]
mx	Mexico	2010–2025 (16 files)	http://www.dgis.salud.gob.mx/contenidos/basesdedatos/da_egresoshosp_gobmx.html	[2026-06-08]
cl	Chile	2001–2025 (25 files)	https://deis.minsal.cl/#datosabiertos	[2026-06-03]
ny	USA, New York State	2009–2024 (16 files)	https://health.data.ny.gov/browse Search phrase: Hospital Inpatient Discharges (SPARCS De-Identified)	[2026-06-08]

Table S3. Release package structure, as released.

Artifact	Role
crosswalk/kr.json	Declarative rules for the Korean source.
crosswalk/br.json	Declarative rules for the Brazilian source.
crosswalk/mx.json	Declarative rules for the Mexican source.
crosswalk/cl.json	Declarative rules for the Chilean source.
crosswalk/ny.json	Declarative rules for the New York State source.
crosswalk/mx_2016.json	One-time pre-step for Mexico 2016 date order normalization; non-idempotent and must be run exactly once before the main crosswalk.
assets/kr_krw.json	Annual KRW/USD exchange-rate factors used to derive EDEC_TRAMT_USD for Korea.
assets/ny_usd.json	Annual factor 1.0 for New York State because the native currency is USD.
assets/kr_age_group.json	Dictionary mapping Korean age-group codes to interval labels.
pipeline/	Shared reader, transformation and derivation functions.
validation/	Scripts for schema checks, value-domain checks, completeness checks, row counts, checksums and aggregate comparisons.

docs/source_access/	Access records with provider name, URL, licence/terms, access date, expected filename pattern and file checksum where available.
logs/	Machine-readable logs from ingestion, harmonization and validation.
examples/	Minimal reproducible examples or synthetic files that demonstrate the crosswalk execution without redistributing restricted microdata.

3. Canonical column model and field availability

The canonical column model contains 14 fields: CNTRY_CD plus 13 substantive fields, using English NHIS-derived names as stable technical targets. It is intentionally lightweight: it links the five sources reproducibly but does not claim full semantic equivalence between all values or replace comprehensive common data models such as OMOP CDM.

Canonical field names are stable technical slots rather than assertions of semantic equivalence, and each field carries a comparability class stating how far cross-source pooling is supported. Class C1 covers fields carrying the same concept, unit and coding system in every source that produces them. Class C2 covers fields carrying the same concept at a different granularity, unit or anchor, where pooling requires an explicit user-side transformation that the declaration names. Class C3 covers source-specific proxies, where the shared name indicates structural position only and pooling is not supported. A class is a property of the pair of sources being combined rather than of the column alone, so it is evaluated over the sources actually being pooled, and the field-level class recorded in Table S4 is the weakest class over the sources producing the field. Classes are declared in `schema/canonical_fields.json`, validated against `schema/canonical_fields.schema.json`, emitted per column in a sidecar metadata file alongside each harmonized output, and enforced by the merge helper, which refuses to pool a C3 field across sources unless the caller passes an explicit override and logs a warning when it does. SUB_SICK shows why the distinction is needed: Chile supplies an external-cause code into that slot, while Mexico excludes its own external-cause field and supplies a genuine second diagnosis. Both decisions are deliberate, they run in opposite directions, and the class assignment records that in the data rather than in a footnote.

Table S4. Canonical fields and normalization notes.

Canonical field	Definition	Output format	Normalization note	Class
CNTRY_CD	Source code added during harmonization	Text: kr, br, mx, cl, ny	constant_columns adds a source-specific value.	C1
STND_Y	Base year or file year	YYYY	Usually from source field; in Mexico added from filename by an external step.	C2
KEY_SEQ	Treatment/discharge record identifier	Record identifier	Event-level identifier where available; not a longitudinal patient ID.	C2
SEX	Sex	1=male, 2=female; source-specific unknown values may remain	Brazil 3->2; Chile HOM-BRE/MUJER->1/2; New York M/F->1/2 and U retained.	C1
AGE_GROUP	Age or age group	Exact age or interval label X-Y / X+	Korea code->label; Brazil/Mexico exact age; Chile/New York interval normalization.	C2

RECU_FR_DT	Admission or treatment start date	YYYYMMDD	Available only where daily date exists.	C2
FORM_CD	Form code or admission/record type	Source code or proxy	Comparable as a source-level proxy only.	C3
DSBJT_CD	Specialty, department or proxy category	Source code or proxy	New York uses APR MDC Code as proxy.	C3
MAIN_SICK	Primary diagnosis	ICD-10, except New York AHRQ CCS/CCSR	Korea truncated code repair; New York retains grouper category.	C3
SUB_SICK	Secondary/coexisting diagnosis or source-specific proxy	ICD-10 where available	Chile DIAG2 is external cause; Mexico only 2010; New York absent.	C3
VSCN	Treatment days or length of stay	Number or source-coded threshold	New York may contain 120+ for long stays.	C3
RECN	Visit/hospitalization days	Number	Available only in Korea and Brazil.	C3
EDEC_TRAMT	Total amount in native currency	Numeric	KRW/BRL/USD depending on source; absent in Mexico/Chile.	C3
EDEC_TRAMT_USD	Total amount in USD	Numeric	Derived for Korea; native for Brazil; factor 1.0 for New York; absent in Mexico/Chile.	C3

Table S5. Semantic profile of the canonical fields whose source concept is not constant across jurisdictions. The class column states the pairs over which pooling is supported.

Canonical field	Source	Source concept as defined by the provider	Coding system or unit	Class
STND_Y	kr	Base year of the extract	YYYY	C2
STND_Y	br	Billing competence period (ANO_CMPT)	YYYY, not the admission year	C2
STND_Y	mx	Year taken from the file name	YYYY, discharge year	C2
STND_Y	cl	Discharge year (ANO_EGRESO)	YYYY	C2
STND_Y	ny	Discharge Year	YYYY	C2
AGE_GROUP	kr	Age-group code	NHIS five-year bands	C2
AGE_GROUP	br	Age at admission (IDADE)	Integer, unit qualifier COD_IDADE not applied	C2
AGE_GROUP	mx	Age at discharge (EDAD)	Integer, unit qualifier CVEEDAD not applied	C2
AGE_GROUP	cl	Age group (GRUPO_EDAD)	Text interval normalized to X-Y / X+	C2
AGE_GROUP	ny	Age Group	Text interval normalized to X-Y / X+	C2
RECU_FR_DT	kr	Treatment start date	YYYYMMDD	C2
RECU_FR_DT	br	Admission date (DT_INTER)	YYYYMMDD	C2
RECU_FR_DT	mx	Admission date	YYYYMMDD	C2
FORM_CD	kr	Claim form code	NHIS form code	C3
FORM_CD	br	AIH record type (IDENT)	SIHSUS record-type code	C3
FORM_CD	mx	Type of service (TIPSERV)	SAEH service catalogue	C3

FORM_CD	ny	Type of Admission	SPARCS admission-type code	C3
DSBJT_CD	kr	Treating department code	NHIS specialty code	C3
DSBJT_CD	br	Specialty (ESPEC)	SIHSUS specialty code	C3
DSBJT_CD	mx	Discharge service (SER-VICIOEGRE)	saeh_CATSERVICIOS catalogue	C3
DSBJT_CD	ny	APR MDC Code	Organ-system category, not a department	C3
MAIN_SICK	kr	Principal diagnosis	ICD-10 (KCD)	C2 among ICD-10 sources
MAIN_SICK	br	Principal diagnosis (DIAG_PRINC)	ICD-10 (CID-10)	C2 among ICD-10 sources
MAIN_SICK	mx	Principal condition (AFECPRIN)	ICD-10 (CIE-10)	C2 among ICD-10 sources
MAIN_SICK	cl	Principal diagnosis (DIAG1)	ICD-10 (CIE-10)	C2 among ICD-10 sources
MAIN_SICK	ny	Grouper category	AHRQ CCS 2009-2017, CCSR 2018-2024	C3
SUB_SICK	kr	Coexisting diagnosis	ICD-10	C2 with br, mx
SUB_SICK	br	Secondary diagnosis (DIAG_SECUN)	ICD-10, empty from 2015 onward	C2 with kr, mx
SUB_SICK	mx	Second condition (SEGUNDA_AFECCION), 2010 only	ICD-10, external cause CAUSAEXT deliberately excluded	C2 with kr, br
SUB_SICK	cl	External cause (DIAG2)	ICD-10 external-cause chapters, not a comorbidity	C3
VSCN	kr	Treatment days billed	Integer days, not length of stay	C3
VSCN	br	Days of stay (DIAS_PERM)	Integer days	C2 with mx, cl, ny
VSCN	mx	Days of stay (DIAS_ESTA)	Integer days	C2 with br, cl, ny
VSCN	cl	Days of stay (DIAS_ESTADA)	Integer days	C2 with br, mx, ny
VSCN	ny	Length of Stay	Integer days, threshold-coded at 120	C2 with br, mx, cl
RECN	kr	Visit or hospitalization days	Integer days	C3
RECN	br	Number of daily rates billed	Integer count, not days of stay	C3
EDEC_TRAMT	kr	Reimbursed claim amount	KRW	C3
EDEC_TRAMT	br	Authorized AIH value (VAL_TOT)	BRL	C3
EDEC_TRAMT	ny	Total Charges, billed rather than incurred	USD	C3
EDEC_TRAMT_USD	kr	Derived from the reimbursed claim amount	USD via annual FRED AEXKOUS rate	C3
EDEC_TRAMT_USD	br	Native US_TOT column published by DATASUS	USD, conversion rate and date not documented by the provider	C3
EDEC_TRAMT_USD	ny	Identity transform of Total Charges	USD, factor 1.0	C3

Table S6. Field availability matrix.

Canonical field	KR	BR	MX	CL	NY
CNTRY_CD*	✓	✓	✓	✓	✓
STND_Y	✓	✓	≈ ¹	✓	✓
KEY_SEQ	✓ ²	✓	✓	—	—
SEX	✓	✓	✓	✓ ⁵	✓ ⁹
AGE_GROUP	✓	≈	≈	≈ ⁶	≈ ⁶
RECU_FR_DT	✓	✓	✓	— ⁷	— ⁷
FORM_CD	✓	≈	≈	—	≈
DSBJT_CD	✓	✓	✓	—	≈ ¹⁰
MAIN_SICK	✓	✓	✓	✓	≈ ¹¹
SUB_SICK	✓	✓ ¹⁴	≈ ³	≈ ⁸	—
VSCN	✓	✓	✓	✓	✓
RECN	✓	✓ ¹³	—	—	—
EDEC_TRAMT	✓	✓	—	—	✓
EDEC_TRAMT_USD	≈ ⁴	✓	—	—	≈ ¹²
Σ SUM	14	14	10 (11 in 2010)	7	10

A baseline field set is defined for each source, while vintage-specific departures from that baseline are documented separately as provider-side schema drift. The Korean record identifier settles the point: it is present in 21 of 22 vintages and absent in 2020, and a source-level declaration could only report a false absence in 21 files or conceal a real one in the twenty-second. Seventeen files deviate from a source-level profile, and each deviation is provider-side schema drift recorded in this matrix rather than a defect in the harmonized output: the Mexican source stops publishing SEGUNDA_AFECCION, the input to SUB_SICK, from the 2011 vintage onward, which accounts for fifteen files, and the Brazilian source does not publish QT_DIARIAS, the input to RECN, in 2006 or 2007, which accounts for two. Structurally absent fields account for the remainder of the gap to the full canonical schema: seven fields across all 25 Chilean vintages, four across 16 New York vintages, three across 16 Mexican vintages and one across two Brazilian vintages. Full canonical conformance as delivered is reached by 39 files, 21 Korean and 18 Brazilian.

Drift affects the content of a column as well as its presence, which is a further reason the declaration cannot be made at source level. The Mexican admission-date column is published in five notations across vintages, the New York monetary column in four, the New York diagnosis column changes grouper family at the 2017 to 2018 boundary, and the Chilean age labels take at least four forms. Each is documented in Table S7 and exercised by the record audit.

Candidates were identified by inspecting national open-data portals, since public discharge microdata are catalogued alongside general government data rather than in a dedicated health channel. The screening criterion was whether a record-level file exposes a diagnosis code, an age or age band and sex in a form that maps deterministically onto the column model. Two exclusion classes dominate. Sources published only as periodic aggregates have no record-level file to harmonize; the Italian national portal is representative, publishing monthly summaries. Sources that publish record-level rows but suppress the diagnosis code cannot support a case definition; the Polish inpatient reporting dataset on dane.gov.pl is representative, publishing rows in which the diagnosis code is masked. A third class covers sources whose record-level files carry too few of the model's fields or carry the clinical field in uncoded form: the Peruvian files provide a year, a diagnosis and a region without usable sex or age, and the Uruguayan files describe the diagnosis in free

text rather than by code. Coarse age banding on its own is not a reason for exclusion, since Chile and New York publish age only as intervals and both are included. All three classes remove a source from consideration entirely rather than making it harmonizable at a cost, which is why the inclusion criterion is stated as compatibility with the column model and why the resulting set is described as five case studies rather than as a sample.

Legend

✓ full field coverage

≈ approximate, proxy, or conditional (see footnote)

— absent (column not produced in output)

Footnotes

* CNTRY_CD added during harmonisation as a constant (kr/br/mx/cl/ny); does not exist in source data.

1. column absent; year is derived from the filename and added in a separate step (MX).

2. absent in the 2020 release (KR).

3. 2010. only (SEGUNDA_AFECCION); column absent in all other vintages (MX).

4. derived from the exchange-rate file assets/kr_krw.json, not a native field (KR).

5. HOMBRE/MUJER normalised to 1/2; 2012 and 2021 already numeric (CL).

6. text interval reduced to X–Y; granularity varies across vintages (CL, NY).

7. no daily date available; year only (CL ANO_EGRESO, NY Discharge Year, both mapped to STND_Y).

8. DIAG2 is an external cause (causa externa), not a secondary diagnosis; frequently empty (CL).

9. M/F normalised to 1/2; U (Unknown) retained (NY).

10. proxy: APR MDC Code (organ-system category), not a ward/department code (NY).

11. AHRQ grouper, not ICD-10: CCS (2009–2017) / CCSR (2018–2024) (NY).

12. native currency = USD; derived as EDEC_TRAMT / 1.0 (assets/ny_usd.json) (NY).

13. absent or completely empty in the 2006 and 2007 releases (BR).

14. completeness is 0.0 from 2015 to 2025; effectively absent in these vintages (BR).

4. Source-to-canonical field mapping and authoritative variable provenance

The table below maps each canonical field to the source column that feeds it. A dash indicates that the source does not produce the field. The mapping is implemented in the corresponding crosswalk JSON files, which are the executable source of truth. Table S8 adds variable-level semantic provenance and identifies the provider documentation used to interpret each source field.

Table S7. Source-to-canonical mapping.

Canonical field	KR	BR	MX	CL	NY
CNTRY_CD	(constant kr)	(constant br)	(constant mx)	(constant cl)	(constant ny)
STND_Y	Base year	ANO_CMPT	(filename)	ANO_EGRESO	Discharge Year
	Treatment/claim				
KEY_SEQ	record sequence identifier	N_AIH	ID	-	-
SEX	Sex code	SEX0	SEX0	SEX0	Gender
AGE_GROUP	Age-group code	IDADE	EDAD	GRUPO_EDAD	Age Group

RECU_FR_DT	Treatment start date	DT_INTER	INGRE	-	-
FORM_CD	Claim form code	IDENT	TIPSERV	-	Type of Admission
DSBJT_CD	Treating department/specialty code	ESPEC	SERVICIOEGRE	-	APR MDC Code
MAIN_SICK	Principal diagnosis code	DIAG_PRINC	AFECPRIN	DIAG1	CCS / CCSR Diagnosis Code
SUB_SICK	Coexisting/secondary diagnosis code	DIAG_SECUND	SEGUNDA_AFEC-CION	DIAG2	-
VSCN	Treatment days billed	DIAS_PERM	DIAS_ESTA	DIAS_ESTA-TADA	Length of Stay
RECN	Visit/hospitalization-day count	QT_DI-ARIAS	-	-	-
EDEC_TRAMT	Adjudicated/reimbursed total health-insurance benefit amount	VAL_TOT	-	-	Total Charges
EDEC_TRAMT_USD	(derived)	US_TOT	-	-	(derived)

To distinguish semantic provenance from executable transformation logic, Table S8 links each source variable used by the crosswalks to the provider-defined meaning and the authoritative provider documentation used to interpret that field. The applicable vintage is stated explicitly where availability or classification changes over time. The crosswalk JSON files remain the executable source of truth for the implemented transformation; the provider documents listed below are the semantic source of truth where a citable field-level specification is available.

Table S8. Authoritative provider documentation and variable-level provenance for source fields used in the crosswalks.

Source	Vintage(s)	Source variable	Provider-defined meaning / role	Canonical field	Doc.
KR	2002–2023	Base year	Base year of the extract.	STND_Y	KR-A*
KR	2002–2019; 2021–2023	Treatment/claim record sequence identifier	Treatment/claim record sequence identifier; event-level identifier, not a longitudinal patient identifier.	KEY_SEQ	KR-A*
KR	2002–2023	Sex code	Sex code.	SEX	KR-A*
KR	2002–2023	Age-group code	Age-group code; represented in the harmonized data as NHIS five-year age bands.	AGE_GROUP	KR-A*

Source	Vintage(s)	Source variable	Provider-defined meaning / role	Canonical field	Doc.
KR	2002–2023	Treatment start date	Treatment start date.	RECU_FR_DT	KR-A*
KR	2002–2023	Claim form code	Claim form code.	FORM_CD	KR-A*
KR	2002–2023	Treating department/specialty code	Treating department/specialty code.	DSBJT_CD	KR-A*
KR	2002–2023	Principal diagnosis code	Principal diagnosis code (KCD/ICD-10 representation in the source).	MAIN_SICK	KR-A*
KR	2002–2023	Coexisting/secondary diagnosis code	Coexisting/secondary diagnosis code.	SUB_SICK	KR-A*
KR	2002–2023	Treatment days billed	Treatment days billed; not assumed to be equivalent to inpatient length of stay.	VSCN	KR-A*
KR	2002–2023	Visit/hospitalization-day count	Visit/hospitalization-day count.	RECN	KR-A*
KR	2002–2023	Adjudicated/reimbursed total health-insurance benefit amount	Adjudicated/reimbursed total health-insurance benefit amount in KRW.	EDEC_TR_AMT	KR-A*
BR	2006–2025	ANO_CMPT	AIH billing competence period; its year component is used as the canonical vintage and is not interpreted as admission year.	STND_Y	BR-A
BR	2006–2025	N_AIH	AIH authorization number / discharge-record identifier.	KEY_SEQ	BR-A
BR	2006–2025	SEXO	Patient sex.	SEX	BR-A
BR	2006–2025	IDADE	Age at admission; the source also carries an age-unit qualifier.	AGE_GROUP	BR-A
BR	2006–2025	DT_INTER	Admission date.	RECU_FR_DT	BR-A
BR	2006–2025	IDENT	AIH record/type identifier.	FORM_CD	BR-A

Source	Vintage(s)	Source variable	Provider-defined meaning / role	Canonical field	Doc.
BR	2006–2025	ESPEC	Specialty code.	DSBJT_CD	BR-A
BR	2006–2025	DIAG_PRINC	Principal diagnosis, coded in CID-10/ICD-10.	MAIN_SICK	BR-A
BR	2006–2025	DIAG_SECUN	Secondary diagnosis, coded in CID-10/ICD-10; effectively empty in the harmonized vintages from 2015 onward.	SUB_SICK	BR-A
BR	2006–2025	DIAS_PERM	Days of stay.	VSCN	BR-A
BR	2008–2025	QT_DIARIAS	Number of daily rates billed; absent or completely empty in the 2006 and 2007 releases.	RECN	BR-A
BR	2006–2025	VAL_TOT	Total authorized AIH value in BRL.	EDEC_TRAMT	BR-A
BR	where present	US_TOT	Native USD-total field published by DATASUS; the provider documentation available to us does not state the conversion date/rate used to create it.	EDEC_TRAMT_USD	BR-A
MX	2010–2025	annual file-name	Discharge year derived from the provider's annual file name; no source column is used for this canonical field.	STND_Y	MX-A
MX	2010–2025	ID	Discharge-record identifier.	KEY_SEQ	MX-A
MX	2010–2025	SEXO	Patient sex.	SEX	MX-A
MX	2010–2025	EDAD	Age at discharge; the source also provides an age-unit qualifier.	AGE_GROUP	MX-A
MX	2010–2025	INGRE	Admission date.	RECU_FRDT	MX-A
MX	2010–2025	TIPSERV	Type of service.	FORM_CD	MX-A
MX	2010–2025	SERVICIO-EGRE	Discharge service, interpreted through the SAEH service catalogue.	DSBJT_CD	MX-A
MX	2010–2025	AFECPRIN	Principal condition/diagnosis, coded in CIE-10/ICD-10.	MAIN_SICK	MX-A
MX	2010 only	SE-GUNDA_AFECCION	Second condition/diagnosis; CAUSAEXT is a separate external-cause field and is deliberately not substituted for this variable.	SUB_SICK	MX-A

Source	Vintage(s)	Source variable	Provider-defined meaning / role	Canonical field	Doc.
MX	2010–2025	DIAS_ESTA	Days of stay.	VSCN	MX-A
CL	2001–2025	ANO_EGRES O	Discharge year.	STND_Y	CL-A
CL	2001–2025	SEXO	Patient sex.	SEX	CL-A
CL	2001–2025	GRUPO_EDA D	Age group published by DEIS; interval labels vary across vintages.	AGE_GR OUP	CL-A
CL	2001–2025	DIAG1	Principal diagnosis, coded in CIE-10/ICD-10.	MAIN_SI CK	CL-A
CL	2001–2025	DIAG2	External cause (causa externa), coded in ICD-10; not interpreted as a comorbidity diagnosis.	SUB_SICK	CL-A
CL	2001–2025	DIAS_ES- TADA	Days of stay.	VSCN	CL-A
NY	2009–2024	Discharge Year	Calendar year of discharge.	STND_Y	NY-A
NY	2009–2024	Gender	Patient gender; source-specific Unknown value may be retained.	SEX	NY-A
NY	2009–2024	Age Group	Age reported in grouped intervals rather than as exact age.	AGE_GR OUP	NY-A
NY	2009–2024	Type of Admission	Manner/type of admission to the facility.	FORM_C D	NY-A
NY	2009–2024	APR MDC Code	APR major diagnostic category; an organ-system grouping used only as a proxy for the canonical department/specialty slot.	DSBJT_C D	NY-A
NY	2009–2017	CCS Diagnosis Code	AHRQ Clinical Classifications Software (CCS) diagnosis category; a grouper category rather than a raw diagnosis code.	MAIN_SI CK	NY-A + NY- CCS
NY	2018–2024	CCSR Diagnosis Code	AHRQ Clinical Classifications Software Refined (CCSR) diagnosis category; a grouper category rather than a raw diagnosis code.	MAIN_SI CK	NY-A + NY- CCSR
NY	2009–2024	Length of Stay	Patient days for the stay; long stays are source-coded as a threshold category (120+ in applicable de-identified releases).	VSCN	NY-A
NY	2009–2024	Total Charges	Total billed charges in USD; retained as a charge concept, not interpreted as incurred cost.	EDEC_TR AMT	NY-A

Documentation keys and provenance notes. All URLs below are provider or government documentation resources (access date is 12 August 2026).

KR-A*. National Health Insurance Service (NHIS), provider source dataset/access record used in this study via the Korea Public Data Portal, dataset no. 15007122: <https://www.data.go.kr/data/15007122/fileData.do> (source-access record: 29 May 2026), together with the original provider file headers retained in docs/source_access. A separate public field-level codebook for this withdrawn extract could not be located at revision time; KR-A* therefore documents provider provenance but is not claimed as an independent field-level dictionary.

BR-A. Ministério da Saúde, DATASUS/SIH-SUS. Official SIH technical documentation and RD (Reduced AIH) file description/layout: https://wiki.saude.gov.br/sih/index.php/P%C3%A1gina_principal and <https://siab.datasus.gov.br/DATASUS/index.php?acao=13&area=0901&item=1¬icia=7760>.

MX-A. Secretaría de Salud, Dirección General de Información en Salud (DGIS), Egresos Hospitalarios. Annual Descriptores Bases de Datos for the applicable vintage and official data portal: https://www.dgis.salud.gob.mx/contenidos/basesdedatos/da_egresoshosp_gobmx.html.

CL-A. Ministerio de Salud de Chile, Departamento de Estadísticas e Información de Salud (DEIS), Egresos Hospitalarios. Applicable annual Estructura IEEH / Esquema de Registro: <https://deis.minsal.cl/sistemas/>; historical field structure: https://repositorio-deis.minsal.cl/deis/egresos_2003/Estructura_2003.htm.

NY-A. New York State Department of Health (NYSDOH), SPARCS. Annual de-identified inpatient data dictionaries / Health Data NY metadata and SPARCS Data Access documentation: <https://www.health.ny.gov/statistics/sparcs/access/>. Annual releases provide the applicable field definitions; examples include 2018: <https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/yjgt-tq93> and 2024: <https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/sf4k-39ay>.

NY-CCS. Agency for Healthcare Research and Quality (AHRQ), Healthcare Cost and Utilization Project (HCUP), Clinical Classifications Software (CCS) resources: <https://hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp>; archived CCS/CCSR resources: https://hcup-us.ahrq.gov/toolssoftware/ccsr/ccsr_archive.jsp.

NY-CCSR. AHRQ/HCUP, Clinical Classifications Software Refined (CCSR) for ICD-10-CM Diagnoses: <https://hcup-us.ahrq.gov/toolssoftware/ccsr/dxccsr.jsp>.

Notes. CNTRY_CD is an analyst-added constant and therefore has no provider source variable. Mexico STND_Y is derived from the provider's annual filename and is included above as artifact-level provenance. KR and NY EDEC_TRAMT_USD are derived fields rather than native source variables and are documented by the exchange-rate/identity-transform assets de-scribed elsewhere in this Supplement. Blank or unavailable provider documentation is not replaced by inferred definitions.

5. Crosswalk execution model

Each source is harmonized through a declarative crosswalk executed by a shared pipeline rather than by source-specific scripts. Crosswalk operations follow a fixed order:

column names are normalized before value transformations, derived columns are created before field selection, and final ordering is applied after selection.

Table S9. Pipeline stages.

Stage	Function	Example
read	Read source files in chunks using declared encoding and separator; fall back to robust parsing when needed.	utf-8-sig removes BOM in Mexico; latin-1 safely decodes Chile/New York.
rename	Map source headers and year-specific header variants to canonical names.	DIAG_PRINC or CCS Diagnosis Code -> MAIN_SICK.
transform	Normalize values by regex or dictionary-based replacement.	\$6 679,65 -> 6679.65; HOMBRE -> 1.
convert	Create derived or constant columns.	CNTRY_CD = "kr"; EDEC_TRAMT_USD = EDEC_TRAMT / exchange factor.
select	Keep fields intended for the canonical output and remove source-only fields.	Brazil 113 input columns -> 13 substantive fields plus CNTRY_CD.
order	Apply canonical field order and place CNTRY_CD first; fields not produced by a source are omitted.	CNTRY_CD, STND_Y, KEY_SEQ, ... EDEC_TRAMT_USD.
reindex (optional)	Create a full 14-field rectangular schema across sources by adding missing fields as nulls.	Needed before simple vertical concatenation when the analysis requires all fields present.

The read stage is expanded below, because it is the only stage that can drop a row. The pseudocode covers every recovery branch; the corresponding implementation is in the reader module of the released package, and each branch is pinned by a unit test.

```

for each source file:
    scan pass: tokenize every row and record its field count
    if every row matches the header field count: parse with the
fast reader and stop
    otherwise parse row by row, with k = header field count and
n = observed field count:
        if n = k: accept the row unchanged
        else if n > k and the trailing (n - k) fields are all
empty: trim to k, accept, count as repaired
        else: write the row to quarantine with its line number,
n, k and a reason code, and exclude it
    write the ledger: provider lines, parsed rows, repaired
rows, quarantined rows, output records

```

No branch realigns a row. Trimming removes only empty trailing fields, so it cannot change the column into which any retained value falls, and every other deviation is quarantined rather than repaired. Reason codes distinguish too few fields, too many fields, and a row collapsed to a single field. Input and output field counts are recorded for every quarantined row, and the ledger reconciles the five row counts above for all 99 files.

Idempotence is a design requirement. rename lists include both source and canonical names, and value transformations target only unnormalized forms. Two distinct properties are involved and are reported separately. Crosswalk idempotence means that re-running the main crosswalk on its own output leaves the file unchanged byte for byte; this is tested for all five sources and is the property the 99/99 figure refers to. End-to-end reproducibility means that the whole chain from the provider file, with one-time pre-steps executed exactly once, yields the same result. The Mexico 2016 pre-step is a pre-step and not a crosswalk, and is one-time by design. Checksums are computed on the uncompressed

content of the output file, with source row order preserved and no sorting applied, and with the number format fixed by the decimals setting of the arithmetic rule; all other values are carried as text and never pass through a floating-point representation. Re-reading a harmonized file must use that file's output delimiter rather than the source delimiter: the Chilean crosswalk reads a semicolon-delimited source and writes a comma-delimited output, so a naive re-run that keeps the source delimiter parses each line as a single column and destroys the file. The idempotence check substitutes the delimiter before re-running, and both behaviours are pinned by tests.

The convert step is executed by the pipeline engine (pipeline/derive.py): the `apply_constants` function introduces constant columns, such as `CNTRY_CD`, between the transform and select stages. Merging harmonized sources without the optional reindex step requires column alignment by name (e.g., via `pandas.concat()`), which inserts missing values where a source does not produce a given field. Concatenating by position would misalign the data.

6. Source-specific preprocessing notes

6.1. Republic of Korea (*kr*)

The Korean files are encoded in EUC-KR; fields retained in the canonical model remain stable after decoding. Headers are occasionally heterogeneous across annual volumes: the 2021 file contains a dual header (Korean and English), the duplicate of which is removed prior to harmonization. The 2017 file has an exclusively English header, handled via year-specific rename variants. A portion of the diagnosis codes appears as truncated prefixes (e.g., `J_` or `U_`); from crosswalk version 1.1 both the underscore form and the asterisk form retain the chapter letter alone (`J_` → `J`), so the output asserts exactly what the source records. The rules target only the truncated forms and leave valid codes untouched. The full input schema is broader than the model (approximately 19 fields). Certain columns are mapped but subsequently discarded during the select stage: the regional field, the insured individual identifier (`IDV_ID`), cost components, the total number of prescription days, and the data snapshot date. The record counts in the 2020 and 2022 files are suspiciously low and are treated as source anomalies pending clarification.

6.2. Brazil (*br*)

The Brazilian files have an extensive input schema (113 columns), from which only fields pertinent to the canonical model are retained. Fixed-width padding is removed by trimming rules before subsequent transformations. The value `SEX=3` is normalized to 2 per the crosswalk, whereas `SUB_SICK="0000"` is preserved as a source-level indicator denoting the absence of a comorbidity, rather than being coerced into a missing value. The `US_TOT` variable, where present, is used as the USD cost field. The mortality field (`MORTE`) is not part of the canonical model and is excluded from the final output.

6.3. Mexico (*mx*)

Mexican files have inconsistent naming conventions for both filenames and extensions; before harmonization, all annual files are renamed to the `EGRESOS_YYYY.csv` pattern (singular to plural, `.txt` to `.csv`). Files encoded in UTF-8 with BOM are parsed using `utf-8-sig`. The 2017 file uses a pipe delimiter and requires separate handling; the 2016 file stores dates in `mm/dd/yyyy` format and requires a one-off pre-processing step (`mx_2016.json`). The `SEGUNDA_AFECCION` field is present exclusively in the 2010 volume; subsequent files do not produce the `SUB_SICK` variable. Sentinel values (`NULL`, `-1`,

8888, and 999) are retained and documented, rather than being excluded during harmonization. FORM_CD and DSBJT_CD are proxy variables: DSBJT_CD is derived from SERVICIOEGRE using the saeh_CATSERVICIOS catalog and is not directly comparable to equivalent fields in other sources.

Table S10. Mexico filename and container normalization.

Year(s)	Input state	Action	Target state
2010	EGRESOS_2010.c sv	No filename change	EGRESOS_2010.c sv
2011- 2015	EGRESO_YYYY.c sv	Rename EGRESO_ -> EGRESOS_	EGRESOS_YYYY. csv
2016	EGRESOS_2016.c sv	Filename unchanged; date pre-step required	EGRESOS_2016.c sv
2017	EGRESOS_2017.c sv	Filename unchanged; pipe separator handled in content preprocessing	EGRESOS_2017.c sv
2018	EGRESO_2018.cs v	Rename EGRESO_ -> EGRESOS_	EGRESOS_2018.c sv
2019- 2022	EGRESOS_YYYY. txt	Convert container extension .txt -> .csv	EGRESOS_YYYY. csv
2023- 2024	EGRESOS_YYYY. csv	No filename change	EGRESOS_YYYY. csv
2025	EGRESO_2025.txt	Rename and convert extension	EGRESOS_2025.c sv

6.4. Chile (cl)

The Chilean source files are encoded in Latin-1 and are similarly heterogeneous in naming. The open data files used three prefix variants: EGRE_DATOS_ABIERTOS_YYYY for the majority of annual volumes, a singular variant EGR_DATOS_ABIERTO_YYYY, and EGRESOS_YYYY for specific years. All files were standardized to the EGRE_DATOS_ABIERTOS_YYYY.csv pattern. The DIAG2 field denotes an external cause rather than a standard comorbidity diagnosis; this distinction is preserved as a semantic limitation. The 2022 file parses without loss in the released version, with equal row counts on input and output and no quarantined rows. The discharge status field (CONDICION_EGRESO) is not part of the canonical model.

6.5. New York State (ny)

The New York source files are encoded in Latin-1 and show intra-file heterogeneity. Identical concepts have different headers across annual files (e.g., Health Service Area versus Hospital Service Area, Facility ID versus Permanent Facility Id, Source of Payment versus Payment Typology), mapped by name. Diagnoses shift from the CCS classification (2009–2017) to CCSR (2018–2024). The de-identified SPARCS files lack raw ICD-10 codes; consequently, the MAIN_SICK field stores the AHRQ grouper category rather than the raw code. Mapping CCS/CCSR categories to raw ICD-10 requires separate HCUP cross-walks and covers only approximately 47% of non-US codes directly; the diagnosis is therefore retained at the grouper level as a proxy. The Total Charges field is retained and normalized as a numerical USD value; Total Costs is deliberately excluded from the current canonical model.

7. Value normalization rules

Table S11. Value-level normalization rules.

Domain	Rule summary
Diagnosis	ICD-10 in KR/BR/MX/CL; AHRQ CCS/CCSR in NY. Korean truncated codes matching $^([A-Z])_$ or $^([A-Z])\{3}$ retain the chapter letter alone from cross-walk version 1.1; a bare letter is deliberately not a valid ICD-10 code and is counted separately in validation. New York is not forced into ICD-10.
Sex	Target representation is 1=male, 2=female where possible. Brazil 3->2, Chile HOM-BRE/MUJER->1/2, New York M/F->1/2. Mexico 9 and New York U remain as source-specific unknown values.
Age	Korea age-group codes are mapped to interval labels. Brazil and Mexico exact age values are retained. Chile and New York text intervals are normalized to X-Y or X+ labels. No global rebinning is applied by the crosswalk.
Dates	Daily dates are normalized to YYYYMMDD where available. Chile and New York provide only year-level information. Mexico 2016 requires a one-time date-order pre-step.
Length of stay	VSCN stores treatment days or length of stay as available. New York values may include threshold-coded long stays such as 120+.
Costs	EDEC_TRAMT_USD is derived for Korea using the annual FRED AEXKOUS KRW/USD series; Brazil retains the provider-supplied US_TOT field, and New York uses an identity factor of 1.0 because its native currency is USD. Mexico and Chile do not produce cost fields.
Missing/sentinel values	Source-specific sentinel values are documented but not harmonized into a universal missing-value convention. This avoids introducing unverifiable semantics but requires users to handle sentinels explicitly.
Column ordering	order_columns places CNTRY_CD first and then follows the canonical field order. Fields absent in a source are omitted unless the optional reindex step is run.

8. Technical validation protocol

The technical validation protocol verifies whether the harmonization pipeline behaves as declared. The protocol was run on each source file; results are summarized by source, year and file. The counts in the manuscript tables are taken from the validation logs of the run recorded in RELEASE.json.

Table S12. Validation protocol.

Check	Description	Level	Recommended output
Schema conformity	Verify field set, field order and source-specific availability against the canonical model and field availability matrix. Reported separately as source-profile conformance (the declared field set for that source and vintage) and full canonical conformance (all 14 fields present after the optional reindex step).	File	N compliant / N tested; list deviations.
Row preservation	Compare input and output row counts after parser recovery and logged skips.	File	Input rows, output rows, recovered rows, skipped rows.
Encoding stability	Confirm successful decoding under the declared encoding and absence of unexpected decode errors.	File	N files decoded / N files tested; exceptions.
Value domains	Check SEX, STND_Y, AGE_GROUP, MAIN_SICK, SUB_SICK, VSCN and cost fields against expected source-specific domains.	Field x source	Percentage valid; residual values.
Completeness	Compute non-empty values and documented sentinel counts for each field by source and year.	Field x year	Completeness matrix and sentinel counts.

Plausibility	Assess non-negative length of stay, age range where exact age exists, expected year ranges and continuity of record counts.	Source x year	Flagged anomalies and trend breaks.
Temporal consistency	Verify that the admission or treatment start date (RECU_FR_DT) aligns with the file vintage year (STND_Y) in sources where a daily date is available. Assessable on the 58 files carrying a daily date (kr 22, br 20, mx 16); cl and ny publish year-level information only and are excluded from the denominator by definition. Reported at record level, because deviations arise from the vintage definition and affect a share of records in every Brazilian and Mexican file rather than a few files.	File	Percentage of temporally consistent records; list of justified deviations
Idempotence	Run the main pipeline again on harmonized output and compare checksums.	File	Checksum match: yes/no.
Aggregate comparability	Compare annual discharge counts and length-of-stay summaries with public aggregate comparators where available.	Source x year	Trend alignment and documented definition differences.
Tracer condition	Extract ICD-10 I63 from the four raw-ICD sources and summarize counts, age, sex and length of stay.	Source x year	Successful extraction and descriptive aggregates.

Comparison against an independent aggregate is carried out for New York only, and the scope rule is stated before the result. Such a comparison isolates the behaviour of the pipeline only where the provider publishes an aggregate derived from the same file that we process. New York meets that condition: the state publishes a summary of the same inpatient discharge file, reporting discharge counts and average length of stay by facility, insurance, sex, discharge status and age group. Brazil publishes counts by billing competence period, Korea by the settlement date of a reviewed claim, and Chile and Mexico through system-level tabulations whose institutional scope is not identical to the record-level release, so a discrepancy against any of those would measure a difference of definition rather than a property of the harmonization. Several of them are also published only through interactive online tabulators rather than as downloadable series, which makes the comparison itself hard to reproduce.

Two properties of the published summary have to be handled explicitly, and both are instances of the defects the framework is designed to expose. Discharge counts above one thousand use a non-breaking space as the thousands separator in 9,366 rows, and 849 rows carry the threshold string in place of a numeric average length of stay. Parsing the file without addressing either understates the total by roughly 40 percent. The comparison reported in the main article treats the threshold string as its lower bound and normalizes the separator before summing.

Value-domain conformance is assessed only where a permitted domain can be declared from an authoritative source. The record identifier, the form code and the department code are provider code lists that we do not hold; inventing a domain for them would appeal to a standard that cannot be cited, so they are declared as carrying no domain, their values are counted in a separate category, and both the values and the corresponding cells are excluded from the denominator rather than counted as conforming. This excludes 205 cells and 1,522,653,850 values and leaves 864 assessable cells. The distinction is what makes the metric informative: a field counted as conforming by default cannot register anything, whereas a field declared as unassessed states plainly what the metric does not cover.

Table S13. Data completeness matrix.

	CNTRY_CD	STND_Y	KEY_SEQ	SEX	AGE_GROUP	RECU_FR_DT	FORM_CD	DSBJT_CD	MAIN_SICK	SUB_SICK	VSCN	REC_N	EDEC_TRAMT	EDEC_TRAMT_USD
br_2006	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9978	0.1285	1.0		1.0	1.0
br_2007	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.1256	1.0		1.0	1.0
br_2008	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.1044	1.0	1.0	1.0	1.0
br_2009	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.112	1.0	0.9996	1.0	1.0
br_2010	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.1255	1.0	1.0	1.0	1.0
br_2011	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.1349	1.0	1.0	1.0	1.0
br_2012	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.1419	1.0	1.0	1.0	1.0
br_2013	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.1495	1.0	1.0	1.0	1.0
br_2014	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.1599	1.0	1.0	1.0	1.0
br_2015	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0
br_2016	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0
br_2017	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0
br_2018	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0
br_2019	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0
br_2020	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0
br_2021	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0
br_2022	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0
br_2023	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0
br_2024	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0
br_2025	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0
cl_2001	1.0	1.0		1.0	1.0				1.0	0.0735	1.0			
cl_2002	1.0	1.0		1.0	1.0				1.0	0.0859	1.0			
cl_2003	1.0	1.0		1.0	1.0				1.0	0.0733	1.0			
cl_2004	1.0	1.0		1.0	1.0				1.0	0.0854	1.0			
cl_2005	1.0	1.0		1.0	1.0				1.0	0.0941	1.0			
cl_2006	1.0	1.0		1.0	1.0				1.0	0.0966	1.0			
cl_2007	1.0	1.0		1.0	1.0				1.0	0.0936	1.0			
cl_2008	1.0	1.0		1.0	1.0				1.0	0.0914	1.0			
cl_2009	1.0	1.0		1.0	1.0				1.0	0.0931	1.0			
cl_2010	1.0	1.0		1.0	1.0				1.0	0.0941	1.0			
cl_2011	1.0	1.0		1.0	1.0				1.0	0.093	1.0			
cl_2012	1.0	1.0		1.0	1.0				1.0	0.0937	1.0			
cl_2013	1.0	1.0		1.0	1.0				1.0	0.0917	1.0			
cl_2014	1.0	1.0		1.0	1.0				1.0	0.0911	1.0			
cl_2015	1.0	1.0		1.0	1.0				1.0	0.0919	1.0			
cl_2016	1.0	1.0		1.0	1.0				1.0	0.0929	1.0			
cl_2017	1.0	1.0		1.0	1.0				1.0	0.0944	1.0			
cl_2018	1.0	1.0		1.0	1.0				1.0	0.0926	1.0			
cl_2019	1.0	1.0		1.0	1.0				1.0	0.0955	1.0			
cl_2020	1.0	1.0		1.0	1.0				1.0	0.0992	1.0			
cl_2021	1.0	1.0		1.0	1.0				1.0	0.0944	1.0			
cl_2022	1.0	1.0		1.0	1.0				1.0	0.0941	1.0			
cl_2023	1.0	1.0		1.0	1.0				1.0	0.096	1.0			
cl_2024	1.0	1.0		1.0	1.0				1.0	0.0964	1.0			
cl_2025	1.0	1.0		1.0	1.0				1.0	0.0965	1.0			

kr_2002	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.6176	1.0	1.0	1.0	1.0
kr_2003	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.6213	1.0	1.0	1.0	1.0
kr_2004	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.641	1.0	1.0	1.0	1.0
kr_2005	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.6513	1.0	1.0	1.0	1.0
kr_2006	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.6384	1.0	1.0	1.0	1.0
kr_2007	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.6347	1.0	1.0	1.0	1.0
kr_2008	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.6486	1.0	1.0	1.0	1.0
kr_2009	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.6756	1.0	1.0	1.0	1.0
kr_2010	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.6937	1.0	1.0	1.0	1.0
kr_2011	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.7052	1.0	1.0	1.0	1.0
kr_2012	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.7177	1.0	1.0	1.0	1.0
kr_2013	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.7347	1.0	1.0	1.0	1.0
kr_2014	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.7507	1.0	1.0	1.0	1.0
kr_2015	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.7616	1.0	1.0	1.0	1.0
kr_2016	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
kr_2017	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
kr_2018	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
kr_2019	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
kr_2020	1.0	1.0		1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
kr_2021	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
kr_2022	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
kr_2023	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
mx_2010	1.0	1.0	1.0	1.0	0.9994	1.0	1.0	0.997	1.0	1.0	1.0			
mx_2011	1.0	1.0	1.0	1.0	0.9999	1.0	1.0	0.9977	1.0		1.0			
mx_2012	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9979	1.0		1.0			
mx_2013	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9974	1.0		1.0			
mx_2014	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9975	1.0		1.0			
mx_2015	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0028	1.0		1.0			
mx_2016	1.0	1.0	0.9999	0.9999	0.9999	1.0	0.9999	0.9999	0.9999		0.9999			
mx_2017	1.0	1.0	1.0	1.0	0.9956	0.9986	1.0	0.9807	1.0		1.0			
mx_2018	1.0	1.0	1.0	1.0	0.9999	1.0	1.0	1.0	1.0		1.0			
mx_2019	1.0	1.0	1.0	1.0	0.9999	1.0	1.0	1.0	1.0		1.0			
mx_2020	1.0	1.0	1.0	1.0	0.9991	1.0	1.0	1.0	1.0		1.0			
mx_2021	1.0	1.0	1.0	1.0	0.9973	1.0	1.0	0.9949	1.0		1.0			
mx_2022	1.0	1.0	1.0	1.0	0.9963	1.0	1.0	0.9963	1.0		1.0			
mx_2023	1.0	1.0	1.0	1.0	0.9971	1.0	1.0	0.9943	1.0		1.0			
mx_2024	1.0	1.0	1.0	1.0	0.9986	1.0	1.0	0.9938	1.0		1.0			
mx_2025	1.0	1.0	1.0	1.0	0.9992	1.0	1.0	0.9945	1.0		1.0			
ny_2009	1.0	1.0		1.0	1.0		1.0	1.0	0.9992		1.0		1.0	1.0
ny_2010	1.0	1.0		1.0	1.0		1.0	1.0	0.9991		1.0		1.0	1.0
ny_2011	1.0	1.0		1.0	1.0		1.0	1.0	1.0		1.0		1.0	1.0
ny_2012	1.0	1.0		1.0	1.0		1.0	1.0	1.0		1.0		1.0	1.0
ny_2013	1.0	1.0		1.0	1.0		1.0	1.0	1.0		1.0		1.0	1.0
ny_2014	1.0	1.0		1.0	1.0		1.0	1.0	1.0		1.0		1.0	1.0
ny_2015	1.0	1.0		1.0	1.0		1.0	1.0	1.0		1.0		1.0	1.0
ny_2016	1.0	1.0		1.0	1.0		1.0	1.0	1.0		1.0		1.0	1.0
ny_2017	1.0	1.0		1.0	1.0		1.0	1.0	1.0		1.0		1.0	1.0
ny_2018	1.0	1.0		1.0	1.0		1.0	1.0	1.0		1.0		1.0	1.0
ny_2019	1.0	1.0		1.0	1.0		1.0	1.0	1.0		1.0		1.0	1.0
ny_2020	1.0	1.0		1.0	1.0		1.0	1.0	1.0		1.0		1.0	1.0
ny_2021	1.0	1.0		1.0	1.0		1.0	1.0	1.0		1.0		1.0	1.0
ny_2022	1.0	1.0		1.0	1.0		1.0	1.0	1.0		1.0		1.0	1.0
ny_2023	1.0	1.0		1.0	1.0		1.0	1.0	1.0		1.0		1.0	1.0
ny_2024	1.0	1.0		1.0	1.0		1.0	1.0	1.0		1.0		1.0	1.0

9. Known anomalies and information-loss map

Information loss is defined broadly. It includes row loss, unavailable years, suspected truncation of source files, columns intentionally excluded from the canonical model, fields that cannot be produced by a given source, and semantic loss caused by proxy representations.

Table S14. Information-loss map.

Mechanism	Affected source(s)	Level	Documentation requirement
Rows quarantined by reader	Mexico 2011 and 2012	Rows	Rows that cannot be parsed unambiguously are written to a quarantine file with a reason code. Quarantined rows total 378 and occur only in the Mexican 2011 and 2012 vintages: 169 and 209 rows, respectively, corresponding to 0.0061% and 0.0073% of those provider files. Of these, 174 carry extra fields and 204 carry too few; the latter were previously padded and accepted into the output as complete records. Each damaged row is isolated and the damage is spread evenly through the file, which indicates a content-dependent defect rather than a corrupted export segment. A column-shift test finds every canonical column intact in 339 rows, at least one distorted canonical column in 25, and 14 rows collapsed to a single field. The Chilean 2022 vintage parses without loss in the released files, with 1,597,118 rows on input and on output and no quarantined rows; the figure given for that file in the submitted version is superseded by the ledger.
Rows recovered by bad-line normalization	Mexico and any file with dangling separators	Rows	Recovered rows are counted separately from quarantined rows in the row ledger.
Suspected source-file truncation	Korea 2020 and 2022	Rows	2020 reports 999,999 records; 2022 reports 1,048,575 data records, suggesting spreadsheet row-limit truncation.
Unavailable year	Korea 2024	Rows/year	Provider indicates availability but the file could not be downloaded; exclude from harmonized record totals unless resolved.
Columns intentionally excluded	All sources	Columns	Fields outside the canonical model, such as regional or source-only administrative fields, are not retained in the primary output.
Fields not produced by source	MX, CL, NY and selected KR years	Columns	Output schemas contain 14, 10 or 7 fields before optional reindexing.
Sentinel values retained	Mexico and Brazil	Information	Sentinels remain source-specific and must be handled by users.
Granularity/proxy loss	New York diagnosis; age fields across sources	Information	New York uses CCS/CCSR rather than ICD-10; age granularity differs across sources.

10. Reproducibility and extension guidance

10.1. Minimal reproducibility workflow

The released package includes an executable command sequence; the same sequence is recorded in the commands section of RELEASE.json, with a description of what each step produces. The order of operations is given below.

1. Create or activate the analysis environment and install pinned dependencies.
2. Download source files from the original providers using the source access records.
3. Verify source filenames and checksums where checksums are available.
4. Run file-level normalization for sources that require it, especially Mexico and Chile.
5. Run non-idempotent pre-steps exactly once, especially crosswalk/mx_2016.json.
6. Run the main harmonization pipeline for each source using crosswalk/<source>.json.
7. Optionally run reindex to produce a full 14-field rectangular schema.
8. Run validation scripts and export validation logs, tables and figures.
9. Run checksum comparison for idempotence testing.
10. Deposit code, crosswalks, logs, documentation and permitted derived outputs with a persistent identifier.

10.2. Adding a new jurisdiction

A new jurisdiction can be added without changing the general pipeline if it can be mapped to the canonical field model. The recommended workflow is:

1. Prepare a source access record documenting provider, URL, licence/terms, access date, file pattern, encoding and separator.
2. Profile the input files across all available years to detect schema drift, encoding issues, header variants and row-level parsing defects.
3. Create crosswalk/<new_source>.json with rename, transform, convert, select and order sections.
4. The detailed format of the crosswalk and the full specification of each section (rename, transform, convert, select, order) are documented in the code repository, in crosswalk/README.md.
5. Add helper assets only when the mapping requires a controlled dictionary, exchange-rate table or source-specific derivation factor.
6. Run the validation protocol and document any deviations in the information-loss map.
7. Update the field availability matrix and source-to-canonical mapping tables.

10.3. Adding a new tracer condition

Tracer conditions are implemented outside the core crosswalk and applied to canonical fields produced by the pipeline. The case definition is declared per source and vintage in schema/case_definitions.json, giving the coding system, the matching rule, the year range, whether the definition is strict or approximate, and whether the rule is equivalent to the ICD-10 I63 group; the analysis reads the definition from that file rather than carrying its own copy of the codes. For the four raw-ICD sources the rule is normalized prefix matching on MAIN_SICK against ^I63. For New York State the rule uses grouper categories and is reported separately: CCSR CIR020, Cerebral infarction, is declared equivalent to I63, while CCS 109, Acute cerebrovascular disease, is declared not equivalent, because it also covers intracranial haemorrhage and acute but ill-defined cerebrovascular accidents. The two New York periods are therefore reported as separate rows and neither is added to the strict cohort total.

A second definition, s72_hip_fracture, is declared beside the first and shares its structure, so both cohorts are auditable in the same way. It exercises parts of the mechanism the stroke definition does not. The ICD-10 rule is a set, ^S72[012], rather than a prefix, and the New York rule is an explicit list of grouper codes rather than a single code. The

definition carries a granularity threshold of four characters together with a pattern for records coded at three characters only, which are counted as unassignable and excluded from both the numerator and the denominator instead of being resolved by assumption. It also carries a `reported_beside` declaration for CCSR INJ043, the subsequent-encounter category, which is counted alongside the cohort and never within it, and reporting rules stating that the two New York periods are not summable. The rules are enforced rather than documented: the analysis marks a source whose coding system changes between vintages as not poolable and writes any combined figure to a field labelled as not for quotation, which applies to both definitions.

Grouper category assignments were obtained by sweeping the description columns that the source publishes alongside the codes, yielding 262 CCS and 478 CCSR categories, so that each assignment is traceable to the provider dictionary rather than to external knowledge. Cohort counts per source and vintage, and the pooled statistics with the poolable flag, are released for both definitions.

Mapping correctness is assessed separately from schema and execution checks, because those checks pass equally well when a field is mapped to the wrong concept. A record-level audit samples records per source across vintages, returns to the file the pipeline read, and re-derives every canonical field from the rule written in the source documentation without invoking the crosswalk engine; re-deriving a value with the same code would demonstrate determinism rather than correctness. Record matching is positional and corrected for quarantined rows, which also verifies that row order is preserved. Verdicts distinguish a match, a mismatch, a pass-through where the value arrived already in canonical form and there is nothing to check, and an absent field. A mapping inventory lists all 57 declared mappings across the five sources with columns for the provider document, its version and the position of the relevant definition; those columns are deliberately left empty where we hold no citable reference, because a visible gap is preferable to a plausible-looking invention. The inventory also records which mappings the audit exercised.

The audit established two limits on what it can claim. The Korean input files, and the Brazilian 2006 file, are already in the canonical column layout, including a pre-computed converted-currency column, so they are an intermediate product rather than a provider export. For those files every field is a pass-through, the crosswalk performs only an age-group relabelling, and the currency conversion cannot be recomputed from the available inputs even though its rate source is documented. The Korean 2020 file carries no record identifier at input or output. The audit also recorded a schema drift not described in the provider documentation: the Mexican admission-date column is published in five different formats across vintages, all of which the pipeline normalizes correctly, verified record by record.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.