(Kapitel: BfArM-Daten)

Dieses Kapitel behandelt die Verarbeitung der BfArM-Daten unabhängig von einer konkreten technischen Implementierung.

1 Kode- und Umsteiger-Dateien

Das BfArM stellt für jede neue ICD-10-GM- und OPS-Version Dateien für die Überleitung auf die Vorgänger-Version zum Download zur Verfügung, gebündelt jeweils in einer Zip-Datei: (BfArM, a, Downloads).

Es handelt sich hierbei um CSV formatierte Text-Dateien. CSV steht für "Comma-Separated Values" und ist ein sehr einfaches Format, um Daten zu strukturieren. Es wird ein Satzzeichen verwendet, um den restlichen Text in Spalten zu trennen – laut dem Namen normalerweise ein Komma, aber für die BfArM-Dateien wurde Strichpunkt als Trennzeichen gewählt, wahrscheinlich weil die Klassentitel auch Kommata enthalten können. Weitere Informationen zum CSV Datei-Format finden sich hier: (Bonnefoy et al., 2024, Seite 131f).

Vom BfArM werden Kodes als Schlüsselnummern bezeichnet, wenn diese eindeutig sind und einzelne Überleitungen zwischen Kodes werden Umsteiger genannt.

1.1 Kodes / Schlüsselnummern

Hier beispielhaft die ersten sieben Zeilen der Kode-Datei von ICD-10-GM, Version 2024:

```
UNDEF;Undefined
A00;Cholera
A00.0;Cholera durch Vibrio cholerae O:1, Biovar cholerae
A00.1;Cholera durch Vibrio cholerae O:1, Biovar eltor
A00.9;Cholera, nicht näher bezeichnet
A01;Typhus abdominalis und Paratyphus
A01.0;Typhus abdominalis
```

Anmerkungen:

- Ein Strichpunkt = zwei Spalten
  1. Kode
  2. Klassentitel
- Für OPS ist das Format der Kode-Datei identisch.
- Mit Ausnahme des UNDEF-Eintrags in der ersten Zeile ist die Datei alphabetisch nach dem Kode sortiert. UNDEF ist kein ICD-10-GM- oder OPS-Kode, sondern wird für Umsteiger verwendet, um entfernte beziehungsweise neu hinzugefügte Kodes zu kennzeichnen.
- Die Datei enthält nicht-endständige Kodes – im Beispiel oben A00 und A01. Ein Kode ist endständig, wenn er keine Subkategorie hat. (BfArM, b, Kategorie und Kode in der ICD-10-GM)

1.2 Umsteiger / Überleitungen

Im Gegensatz zu den Kodes haben die Umsteiger für ICD-10-GM und OPS unterschiedliche Formate. Hier also zuerst zwei Ausschnitte aus der Umsteiger-Datei für ICD-10-GM, Version 2017 Überleitung auf Version 2016:

```
A00.0;A00.0;A;A
A00.1;A00.1;A;A
A00.9;A00.9;A;A
A01.0;A01.0;A;A
A01.1;A01.1;A;A
---------------------------
U06.0;UNDEF;A;
UNDEF;Z99.0;;
Z99.0;Z99.0;A;
```

Anmerkungen:

- Drei Strichpunkte = vier Spalten
  1. Alter Kode (2016)
  2. Neuer Kode (2017)
  3. Wenn A: automatisch überleitbar von 2016 auf 2017, sonst nicht
  4. Wenn A: automatisch überleitbar von 2017 auf 2016, sonst nicht
- Der obere Abschnitt umfasst die fünf ersten Zeilen der Umsteiger-Datei.
- Der untere Abschnitt enthält beispielhaft zwei Umsteiger mit UNDEF. UNDEF als neuer Kode heißt der alte Kode wurde entfernt. UNDEF als alter Kode heißt der neue Kode wurde hinzugefügt. In diesem Beispiel wurde Z99.0 umbenannt.
- Die Datei ist alphabetisch nach dem alten Kode sortiert und falls dieser bei mehreren Einträgen identisch ist, anschließend nach dem neuen Kode.
- Es sind nur endständige Kodes enthalten.

Dazu im Vergleich ein einzelner Umsteiger aus dem OPS, Version 2024 Überleitung auf Version 2023:

```
1-100;N;1-100;N;A;A
```

Die zusätzlichen Spalten jeweils nach den Kodes speziell für OPS sagen aus, ob Zusatzkennzeichen notwendig sind, siehe dazu auch: (BfArM, c, Kategorie und Kode im OPS).

1.3 “DRY”-Prinzip

“Don’t Repeat Yourself” ist eines der Kardinalprinzipien in der Software-Entwicklung. Obwohl der Grundsatz, Wiederholungen zu vermeiden, wahrscheinlich schon in der Programmierung angewandt wird seit es diesen Beruf gibt, wurde “DRY” erstmals 1999 ausformuliert von (Thomas and Hunt, 2019, Seite 79ff). In der zwanzigjährigen Jubiläumsausgabe verdeutlichen die Autoren, dass es ihnen hierbei nicht nur um das Schreiben von Programmcode geht, sondern vielmehr um die Absichten hinter einem Prozess. Das heißt eine Änderung der Intention einer Software-Lösung sollte nicht mehrere Änderungen an mehreren Stellen nach sich ziehen.

Bei der Integration der BfArM-Daten kann das “DRY”-Prinzip auf zwei Arten angewandt werden.

1. Bezogen auf Kodiersysteme: Alle Funktionen sollten unabhängig davon funktionieren, ob es sich um ICD-10-GM- oder OPS-Daten handelt. Auch die Aufnahme eines zusätzlichen Systems, beispielsweise ATC, sollte möglichst nur Anpassungen erfordern, die durch Abweichungen in der Integration der Daten dieses Systems notwendig sind.
2. Bezogen auf Versionen: Der Prozess der Datenintegration sollte unabhängig von der Version gleich ablaufen und das gleiche Ergebnis liefern, bezogen auf die Datenstruktur. Jede Abweichung zwischen Versionen sollte nur eine möglichst einfach zu implementierende Modifikation des Gesamtprozesses darstellen. Konkret heißt das beim Hinzufügen einer neuen Version, dass an nur einer Stelle die Abweichungen von der Standardversion angegeben werden sollten und der Datenintegrationsprozess danach einmal angestoßen wird. Idealerweise ändert sich bei einer neuen Version nur die Versionsnummer und die Download-URL.

1.4 Standardverfahren und Abweichungen

Im diesem Abschnitt werden alle Abweichungen der ICD-10-GM- und OPS-Versionen von der als Standard gewählten Version 2024 in Tabellen aufgelistet. Konkret gemeinst sind damit: Version, Download-URL, Pfad der Kode- und Umsteiger-Dateien, Sonstiges. Diese Informationen können dann dem Datenintegrationsprozess in einem strukturierten Dateiformat zur Verfügung gestellt werden. §TODO: Verweis auf konkrete Implementation

Die Download-URL der Zip-Dateien setzt sich wie folgt zusammen:

`https://multimedia.gsb.bund.de/BfArM/downloads/klassifikationen/` ...
& für ICD-10-GM: `icd-10/` ...
& für OPS: `ops/` ...
& einen pro Version unterschiedlichen Teil, siehe URL-Eintrag in den Tabellen.
Die URL dient damit auch als “Single Source of Truth” (Bonnefoy et al., 2024, Seite 257).

Die Kode- und Umsteiger-Dateien sind in einem Verzeichnis enthalten:

`Klassifikationsdateien`
Wenn die Tabellen einen Verzeichnis-Eintrag enthalten, dann wird dieser vorangestellt.
Zum Beispiel für ICD-10-GM Version 2021:
`icd10gm2021syst-ueberl-20201111/Klassifikationsdateien`

Der Pfad der Kode-Datei lautet:

Verzeichnis ...
& für ICD-10-GM: `icd10gm` ...
& für OPS: `ops` ...
& die Version ...
& `syst.txt`
Also zum Beispiel: `Klassifikationsdateien/icd10gm2024syst.txt`
Wenn die Tabellen einen Kodes-Eintrag enthalten, wird dieser stattdessen verwendet.

Das gleiche gilt für die Umsteiger-Datei, nur dass der Dateiname normalerweise so ist:

für ICD-10-GM:	`icd10gm` ...
für OPS:	`ops` ...
& Version	...
& `syst_umsteiger_`	...
& Vorgänger-Version	...
& `_`	...
& Version	...
& `.txt`	Zum Beispiel: `ops2024syst_umsteiger_2023_2024.txt`

1.4.1 Sonstige Abweichungen

Eine kurze Erklärung der unter "Sonstiges" gelisteten Abweichungen:

- Vorab-Version
  Diese Version hat noch keine Seite für die Kode-Suche.
- Zip-Unterdatei
  Die Zip-Datei der 2022 Versionen enthielt weitere Zip-Dateien. Vorher wurden alle Dateien zu einer Versionen nach Verwendungszweck nur in Unterverzeichnisse gegliedert, weswegen die gebündelte Zip-Datei insgesamt relativ groß wurde. Ab 2023 werden die Zip-Unterdateien separat zum Download angeboten.
- ISO-8859-1
  Vor 2009 waren Dateien in ISO-8859-1 kodiert, auch Latin-1 genannt, statt UTF-8.
- Punkt-Strich-Notation, Kreuz-Stern-System
  Die Kodes älterer ICD-10-GM Versionen hatten Sonderzeichen gemäß (BfArM, b).
- 6-Spalten-Umsteiger
  Umsteiger älterer ICD-10-GM Versionen enthielten Informationen zur Mehrfachkodierung.
- Nicht endständige Umsteiger
  Im Gegensatz zu allen anderen Überleitungen sind die Umsteiger-Einträge für ICD-10-GM 2.0 auf 1.3. auch für nicht-endständige Kodes enthalten.
- None statt UNDEF
  Von OPS Version 2009 bis 2004 wurde statt UNDEF der Bezeichner "None" verwendet.
- KOMBI-Kode
  OPS Versionen 2.1 und 2.0 enthalten in der Kodes-Datei einen zusätzlichen Eintrag: KOMBI, "Kombinationsschlüsselnummer erforderlich".
- 6-Spalten-Umsteiger (altes Format), 5-Spalten-Umsteiger, 4-Spalten-Umsteiger
  Die Umsteiger der OPS-Versionen von 2009 bis 2005 waren anders formatiert, weil 2005 die Informationen bezüglich Zusatzkennzeichen hinzukamen und bis 2009 die Spalten unterschiedlich angeordnet waren als in allen neueren OPS Versionen.
- 6-Spalten-Umsteiger (ursprüngliches Format)
  Die Umsteiger für OPS Version 2.1 enthielen zusätzliche Spalten wegen Mehrfachverschlüsselung wie die älteren ICD-10-GM Versionen.
- 3-Spalten-Umsteiger
  OPS Version 2.0 zeigte mit nur einer Spalte an, ob automatische Überleitungen möglich sind.

- Keine Überleitung
  Aus der ältesten Version, die Überleitungen enthält, wird zusätzlich die Kodes-Datei für die Vorgänger-Versionen verarbeitet.

§TODO: Verweise auf 1. Vorab-Version (Frontend), 2. Kodierung, 3. Preprocessing.

1.4.2 ICD-10-GM Versionen

Version	Abweichungen zwischen den Versionen	
2025	URL	`version2025-vorab/icd10gm2025syst-ueberl-vorab.zip`
	Kodes	`Klassifikationsdateien/icd10gm2025syst_vorab.txt`
	Umsteiger	`Klassifikationsdateien/icd10gm2025syst_umsteiger_2024_2025_vorab.txt`
	Sonstiges	• Vorab-Version
2024	URL	`version2024/icd10gm2024syst-ueberl.zip`
	Umsteiger	`Klassifikationsdateien/icd10gm2024syst_umsteiger_2023_20221206_2024.txt`
2023	URL	`version2023/icd10gm2023syst-ueberl_20221206.zip`
	Kodes	`Klassifikationsdateien/icd10gm2023syst_20221206.txt`
	Umsteiger	`Klassifikationsdateien/icd10gm2023syst_umsteiger_2022_2023_20221206.txt`
2022	URL	`vorgaenger/icd10gm2022.zip`
	Sonstiges	• Zip-Unterdatei: `icd10gm2022syst-ueberl.zip`
2021	URL	`vorgaenger/icd10gm2021.zip`
	Verzeichnis	`icd10gm2021syst-ueberl-20201111`
2020	URL	`vorgaenger/icd10gm2020.zip`
	Verzeichnis	`icd10gm2020syst-ueberl`
2019	URL	`vorgaenger/icd10gm2019.zip`
	Verzeichnis	`icd10gm2019syst-ueberl`
2018	URL	`vorgaenger/icd10gm2018.zip`
	Verzeichnis	`x1gut2018`
2017	URL	`vorgaenger/icd10gm2017.zip`
	Verzeichnis	`x1gut2017`
2016	URL	`vorgaenger/icd10gm2016.zip`
	Verzeichnis	`x1gut2016`
2015	URL	`vorgaenger/icd10gm2015.zip`
	Verzeichnis	`x1gut2015`
2014	URL	`vorgaenger/icd10gm2014.zip`
	Verzeichnis	`x1gua2014`
2013	URL	`vorgaenger/icd10gm2013.zip`
	Verzeichnis	`x1gua2013`
2012	URL	`vorgaenger/icd10gm2012.zip`

	Kodes	x1ueb2011_2012/Klassifikationsdateien/ icd10gmsyst2012.txt
	Umsteiger	x1ueb2011_2012/Klassifikationsdateien/ umsteiger_icd10gmsyst2011_icd10gmsyst2012.txt
2011	URL	vorgaenger/icd10gm2011.zip
	Kodes	x1ueb2010_2011/Klassifikationsdateien/ icd10gmsyst2011.txt
	Umsteiger	x1ueb2010_2011/Klassifikationsdateien/ umsteiger_icd10gmsyst2010_icd10gmsyst2011.txt
2010	URL	vorgaenger/icd10gm2010.zip
	Kodes	x1ueb2009_2010/Klassifikationsdateien/ icd10gmsyst2010.txt
	Umsteiger	x1ueb2009_2010/Klassifikationsdateien/ umsteiger_icd10gmsyst2009_icd10gmsyst2010.txt
2009	URL	vorgaenger/icd10gm2009.zip
	Kodes	x1ueb2008_2009/Klassifikationsdateien/ icd10gmsyst2009.txt
	Umsteiger	x1ueb2008_2009/Klassifikationsdateien/ umsteiger_icd10gmsyst2008_icd10gmsyst2009.txt
2008	URL	vorgaenger/icd10gm2008.zip
	Kodes	x1ueb2007_2008/Klassifikationsdateien/ icd10v2008.txt
	Umsteiger	x1ueb2007_2008/Klassifikationsdateien/ umsteiger20072008.txt
	Sonstiges	• ISO-8859-1
2007	URL	vorgaenger/icd10gm2007.zip
	Kodes	x1ueb2006_2007/Klassifikationsdateien/ ICD10V2007.txt
	Umsteiger	x1ueb2006_2007/Klassifikationsdateien/ Umsteiger.txt
	Sonstiges	• ISO-8859-1
2006	URL	vorgaenger/icd10gm2006.zip
	Kodes	x1ueb2005_2006/ICD10V2006.txt
	Umsteiger	x1ueb2005_2006/umsteiger.txt
	Sonstiges	• ISO-8859-1
2005	URL	vorgaenger/icd10gm2005.zip
	Kodes	x1ueb2004_2005/ICD10V2005.txt
	Umsteiger	x1ueb2004_2005/umsteiger.txt
	Sonstiges	• ISO-8859-1
2004	URL	vorgaenger/icd10gm2004.zip
	Kodes	x1ueb20_2004/icd10v2004.txt
	Umsteiger	x1ueb20_2004/Umsteiger.txt

	Sonstiges	• ISO-8859-1 • Punkt-Strich-Notation • 6-Spalten-Umsteiger
2.0	URL	`vorgaenger/icd10gm20.zip`
	Kodes	`x1ueb13_20_v11/icd10v20.txt`
	Umsteiger	`x1ueb13_20_v11/Umsteiger.txt`
	Sonstiges	• ISO-8859-1 • Punkt-Strich-Notation • Kreuz-Stern-System • 6-Spalten-Umsteiger • Nicht endständige Umsteiger
1.3	Kodes	`x1ueb13_20_v11/icd10v13.txt`
	Sonstiges	• ISO-8859-1 • Punkt-Strich-Notation • Kreuz-Stern-System • Keine Überleitung

1.4.3 OPS Versionen

Version	**Abweichungen zwischen den Versionen**	
2025	URL	`version2025-vorab/ops2025syst-ueberl-vorab.zip`
	Kodes	`Klassifikationsdateien/ops2025syst_vorab.txt`
	Umsteiger	`Klassifikationsdateien/` `ops2025syst_umsteiger_2024_2025_vorab.txt`
	Sonstiges	• Vorab-Version
2024	URL	`version2024/ops2024syst-ueberl.zip`
2023	URL	`version2023/ops2023syst-ueberl.zip`
2022	URL	`vorgaenger/ops2022.zip`
	Sonstiges	• Zip-Unterdatei: `ops2022syst-ueberl.zip`
2021	URL	`vorgaenger/ops2021.zip`
	Verzeichnis	`ops2021syst-ueberl`
2020	URL	`vorgaenger/ops2020.zip`
	Verzeichnis	`ops2020syst-ueberl`
2019	URL	`vorgaenger/ops2019.zip`
	Verzeichnis	`ops2019syst-ueberl`
2018	URL	`vorgaenger/ops2018.zip`
	Verzeichnis	`p1sut2018`
2017	URL	`vorgaenger/ops2017.zip`
	Verzeichnis	`p1sut2017`
2016	URL	`vorgaenger/ops2016.zip`
	Verzeichnis	`p1sut2016`
2015	URL	`vorgaenger/ops2015.zip`

	Verzeichnis	p1sut2015
2014	URL	vorgaenger/ops2014.zip
	Kodes	p1sua2014-20131104/Klassifikationsdateien/ops2014syst_20131104.txt
	Umsteiger	p1sua2014-20131104/Klassifikationsdateien/ops2014syst_umsteiger_2013_2014_20131104.txt
2013	URL	vorgaenger/ops2013.zip
	Kodes	p1sua2013/Klassifikationsdateien/ops2013syst_20121113.txt
	Umsteiger	p1sua2013/Klassifikationsdateien/ops2013syst_umsteiger_2012_2013_20121109.txt
2012	URL	vorgaenger/ops2012.zip
	Kodes	p1ueb2011_2012/Klassifikationsdateien/opssyst2012.txt
	Umsteiger	p1ueb2011_2012/Klassifikationsdateien/umsteiger_opssyst2011_opssyst2012.txt
2011	URL	vorgaenger/ops2011.zip
	Kodes	p1ueb2010_2011/Klassifikationsdateien/opssyst2011.txt
	Umsteiger	p1ueb2010_2011/Klassifikationsdateien/umsteiger_opssyst2010_opssyst2011.txt
2010	URL	vorgaenger/ops2010.zip
	Kodes	p1ueb2009_2010/Klassifikationsdateien/opssyst2010.txt
	Umsteiger	p1ueb2009_2010/Klassifikationsdateien/umsteiger_opssyst2009_opssyst2010.txt
2009	URL	vorgaenger/ops2009.zip
	Kodes	p1ueb2008_2009/Klassifikationsdateien/opsamtl2009.txt
	Umsteiger	p1ueb2008_2009/Klassifikationsdateien/umsteigeramtl20082009.txt
	Sonstiges	• ISO-8859-1 • None statt UNDEF • 6-Spalten-Umsteiger (altes Format)
2008	URL	vorgaenger/ops2008.zip
	Kodes	ops2008amtl/p1ueb2007_2008/Klassifikationsdateien/opsamtl2008.txt
	Umsteiger	ops2008amtl/p1ueb2007_2008/Klassifikationsdateien/umsteigeramtl20072008.txt
	Sonstiges	• ISO-8859-1 • None statt UNDEF • 6-Spalten-Umsteiger (altes Format)
2007	URL	vorgaenger/ops2007.zip
	Kodes	ops2007amtl/p1ueb2006_2007/Klassifikationsdateien/opsamtl2007.txt
	Umsteiger	ops2007amtl/p1ueb2006_2007/Klassifikationsdateien/UmsteigerAmtlich.txt

	Sonstiges	• ISO-8859-1 • None statt UNDEF • 6-Spalten-Umsteiger (altes Format)
2006	URL	`vorgaenger/ops2006.zip`
	Kodes	`ops2006amtl/p1ueb2005_2006/opsv2006.txt`
	Umsteiger	`ops2006amtl/p1ueb2005_2006/umsteiger.txt`
	Sonstiges	• ISO-8859-1 • None statt UNDEF • 6-Spalten-Umsteiger (altes Format)
2005	URL	`vorgaenger/ops2005.zip`
	Kodes	`ops2005amtl/p1ueb2004_2005_v10/OPS2005.txt`
	Umsteiger	`ops2005amtl/p1ueb2004_2005_v10/umsteiger.txt`
	Sonstiges	• ISO-8859-1 • None statt UNDEF • 5-Spalten-Umsteiger
2004	URL	`vorgaenger/ops2004.zip`
	Kodes	`ops2004amtl/p1ueb21_2004_v10/opsv2004.txt`
	Umsteiger	`ops2004amtl/p1ueb21_2004_v10/Umsteiger.txt`
	Sonstiges	• ISO-8859-1 • 4-Spalten-Umsteiger
2.1	URL	`vorgaenger/ops21.zip`
	Kodes	`ops21amtl/p1ueb20_21_v10/opsv21.txt`
	Umsteiger	`ops21amtl/p1ueb20_21_v10/Umsteiger.txt`
	Sonstiges	• ISO-8859-1 • KOMBI-Kode • 6-Spalten-Umsteiger (ursprüngliches Format)
2.0	URL	`vorgaenger/ops20.zip`
	Kodes	`p1ueb11_20_v11/Opsv20.txt`
	Umsteiger	`p1ueb11_20_v11/Umsteiger.txt`
	Sonstiges	• ISO-8859-1 • KOMBI-Kode • 3-Spalten-Umsteiger
1.1	Kodes	`p1ueb11_20_v11/Opsv11.txt`
	Sonstiges	• ISO-8859-1 • Keine Überleitung

2 Datenintegrationsprozess

Wie erwähnt durchlaufen alle Daten unabhängig von Version und Kodiersystem den gleichen Integrationsprozess. Dieser orientiert sich an dem klassischen "Extract-Transform-Load" Modell, siehe (Bonnefoy et al., 2024, Seite 247ff).

1. *Extract:* Die Daten werden in einem bestimmten Format aus einem Quell-System extrahiert.
2. *Transform:* In einem oder mehreren Prozessen werden die Daten in ein standardisiertes Format

transformiert, was zum Beispiel Bereinigung, Validierung und Imputation beinhalten kann.

3. *Insert:* Die Daten werden in ein Ziel-System integriert, um dort von weiteren Applikationen verwendet zu werden.

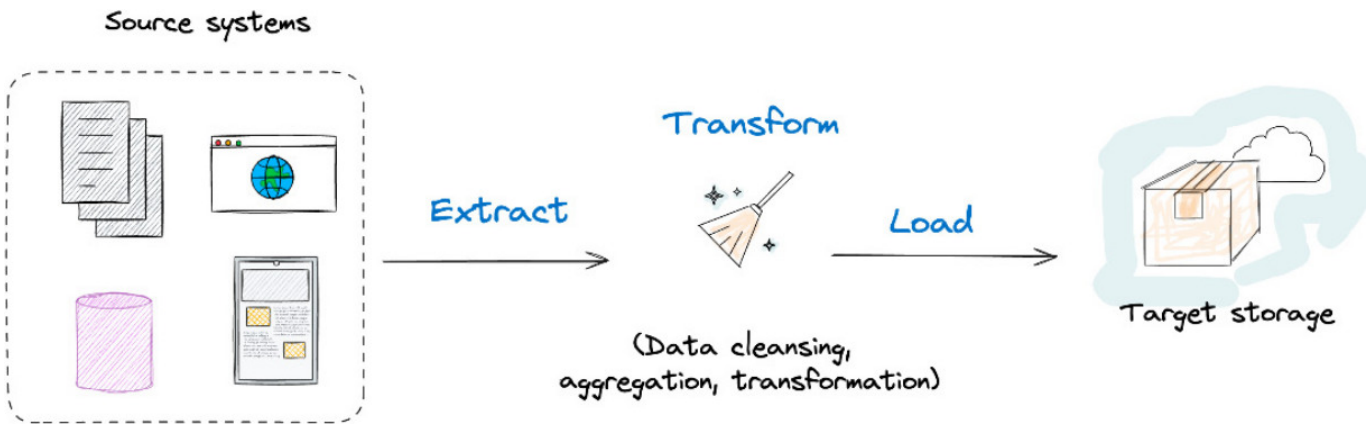


Abbildung 1: ETL-Modell nach (Bonnefoy et al., 2024, Seite 63)

Für die BfArM-Daten sieht der Integrationsprozess konkret so aus:

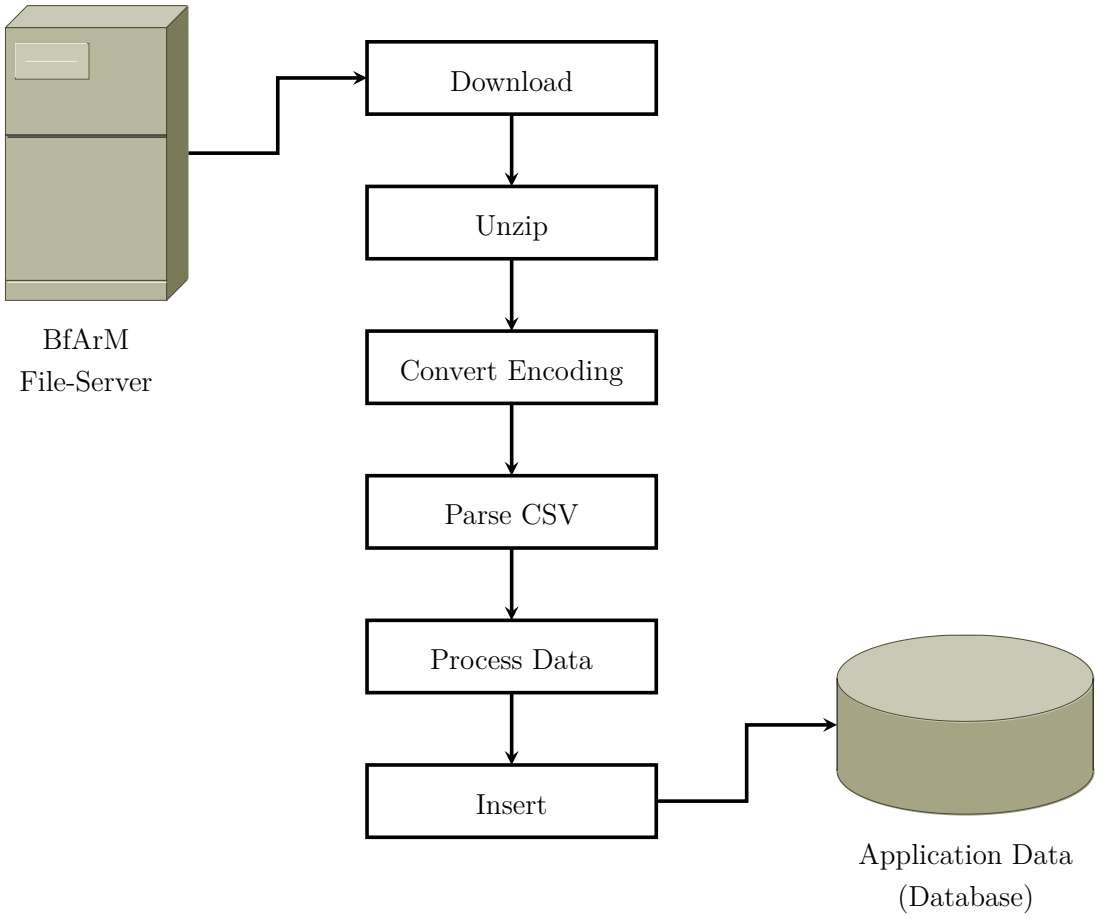


Abbildung 2: BfArM-Datenintegrationsprozess

1. *Download:* Die Zip-Dateien werden heruntergeladen. Alternativ kann geprüft werden, ob die Dateien schon lokal vorhanden sind mit einem bestimmten Pfad, der sich nach Kodiersystem und Versionsnummer immer gleich zusammensetzt, zum Beispiel: `files/icd10gm2024.zip`. Die Download-Funktion sollte die Zip-Dateien ebenfalls unter diesem Pfad abspeichern, falls so gewünscht.

2. *Unzip:* Die Kodes- und Umsteiger-Dateien werden aus der Zip-Datei extrahiert. Normalerweise muss dafür nicht das ganze Archiv in temporäre Dateien entpackt werden – außer eventuell bei den Versionen 2022, weil das Extrahieren verschachtelter Zip-Dateien eher ein Nischenfall ist und nicht unbedingt standardmäßig von Programmiersprachen oder Bibliotheken unterstützt wird.

3. *Convert Encoding:* Die in ISO-8859-1 kodierten Kodes-Dateien müssen in UTF-8 umgewandelt werden. In (Fernández and Manuel, 2022) werden die beiden Zeichenkodierungen genauer erklärt, aber für die BfArM-Daten ist eigentlich nur relevant, dass Umlaute mit unterschiedlichen Werten kodiert sind. Also würde das Einlesen eines in ISO-8859-1 kodierten Umlauts als UTF-8 ein anderes Zeichen als Resultat ergeben. Die Umsteiger-Dateien sind davon nicht betroffen, weil in diesen keine Umlaute enthalten sind.

4. *Parse CSV:* Ein Parser wandelt eine Datei in eine Datenstruktur um; für CSV sollte jede Programmiersprache so eine Funktion standardmäßig zur Verfügung stellen. Für die BfArM-Dateien ist das Ergebnis ein zweidimensionales Array mit zwei Spalten für die Kodes, beziehungsweise drei bis sechs Spalten für die Umsteiger je nach Kodiersystem und Version.

5. *Process Data:* Aufgrund der oben erwähnten Abweichungen ist die Vorverarbeitung der Daten der komplexeste Schritt und wird im nächsten Abschnitt genauer erklärt. Außerdem müssen nicht alle Daten gespeichert werden. Vor allem in Bezug auf die Zip-Dateien ergibt das eine Reduktion der Datenmenge um etwa einen Faktor von zehn.

6. *Insert:* Die bearbeiteten Daten werden für die Verwendung durch Applikationen gespeichert. Zum Beispiel für eine relationale Datenbank werden pro Dateityp, Kodiersystem und Version eine Tabelle angelegt und die Daten in diese geschrieben. Konkret für SQL müssen außerdem die Hochkommata in den Codes-Dateien beachtet werden.

3 Datenvorverarbeitung

"Data Preprocessing" ist ein wichtiger Schritt in Feldern der Informatik wie *Machine Learning* und *Big Data*. In (García et al., 2016) werden mehrere Methoden vorgestellt, wovon folgende in der Verarbeitung der BfArM-Daten zur Verwendung kommen:

1. *Data Cleaning:* Daten werden bereinigt, was sowohl das Korrigieren einzelner Werte, als auch das Entfernen überflüssiger Datensätze beinhaltet. Letzteres wird *Instance Reduction* genannt.

2. *Data Normalization:* Umwandlung der Datensätze auf ein bestimmtes Format.

3. *Data Integration:* Ein Datensatz wird durch zusätzliche Informationen bereichert, beziehungsweise mehrere Informationen werden zu einem Datensatz kombiniert.

4. *Missing values imputation:* Falls Informationen fehlen, müssen die betroffenen Datensätze mit einer bestimmten Logik behandelt werden oder alternativ können Daten durch eine Zufallsfunktion simuliert werden.

Die folgenden Unterabschnitte erklären die Vorverarbeitungsschritte für die BfArM-Daten und beziehen sich damit auf die in 1.4.1 genannten Abweichungen. Die Schritte erfolgen in der gelisteten

Reihenfolge. Obwohl die Daten nach dem CSV-Parsing schon in einer von der Programmiersprache abhängigen Struktur vorliegen, wird zur Erklärung trotzdem noch die Datei-Struktur verwendet.

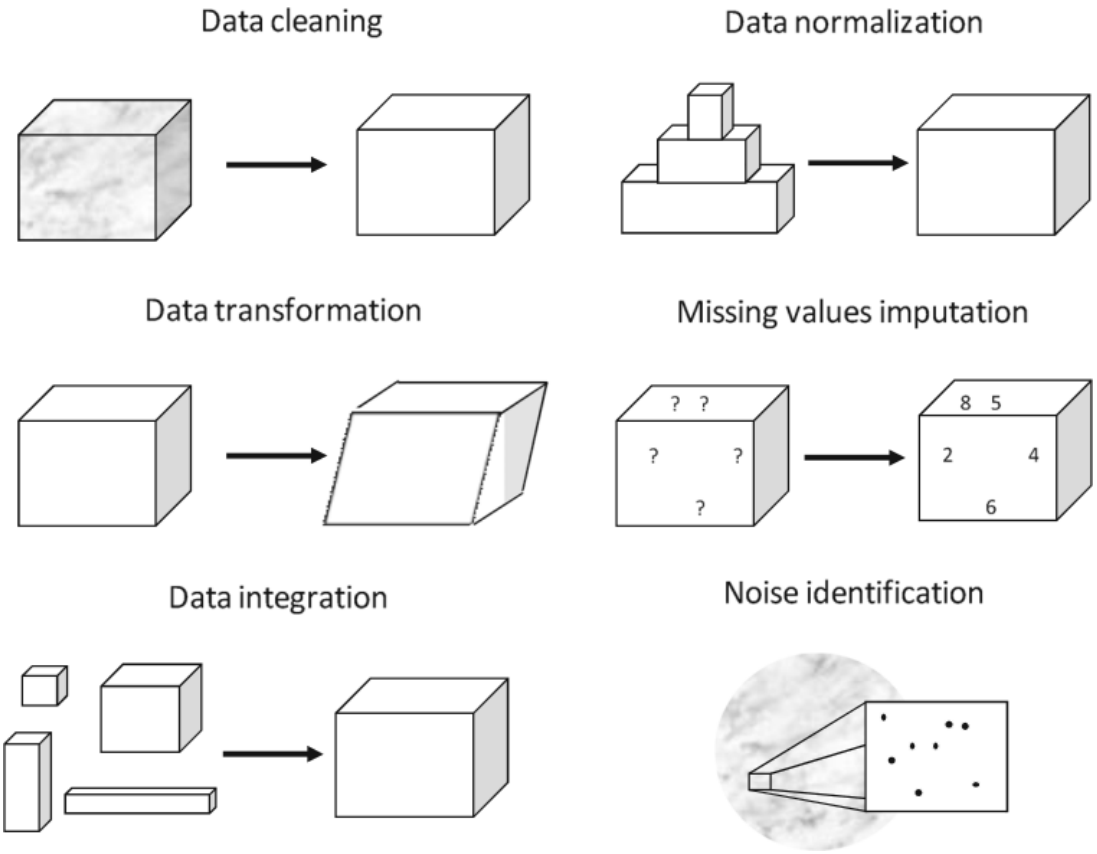


Abbildung 3: “Preprocessing Tasks” aus (García et al., 2016, Seite 4)

3.1 Datennormalisierung

3.1.1 6-Spalten-Umsteiger

Sowohl die ICD-10-GM Versionen 2004 und 2.0, als auch die OPS Version 2.1 beinhalteten Umsteiger in folgendem Format:

```
A00.0;A00.0;A;A;0;UNDEF
-----------------------
1-202;1-202;A;A;;
```

Um diese an das ICD-10-GM Format von 2024 anzupassen, werden die letzten beiden Spalten entfernt.

3.1.2 OPS Umsteiger

Für OPS Versionen ab 2010 sehen die Umsteiger-Einträge so aus:

```
1-100;N;1-100;N;A;A
```

Durch Entfernung der zweiten und vierten Spalte stimmen diese mit den ICD-10-GM Umsteiger Format von 2024 überein.

3.1.3 OPS 6-Spalten-Umsteiger, altes Format

Die OPS Versionen 2009 bis 2006 hatten ebenfalls sechs Spalten für die Umsteiger, aber in einer anderen Reihenfolge:

```
1-100;1-100;N;N;A;A
```

Hier müssen also die dritte und vierte Spalte entfernt werden.

3.1.4 OPS 5-Spalten-Umsteiger

Die Umsteiger von OPS Version 2005 waren in einem ganz eigenen Format geschrieben:

```
5-062.0;5-062.0;N;A;A
5-062.1;5-062.1;N;A;A
5-062.2;5-062.8;J;E;E
5-062.3;5-062.8;J;B;B
```

Hier wird dritte Spalte entfernt und außerdem werden die Sonderformen für automatische Überleitbarkeit von `B` und `E` nach `A` umbenannt.

3.2 Imputation

Für Umsteiger der OPS Version 2.0 gibt es nur drei Spalten:

```
1-208.0;A;1-209.0
-----------------------------
1-208.x;;1-209.4
```

Die zweite Spalte zeigt allein die Überleitbarkeit an. Für die Angleichung an das ICD-10-GM Format von 2024 wird also die zweite Spalte entfernt und gedoppelt angehängt. Aus den beiden Beispielzeilen wird damit:

```
1-208.0;1-209.0;A;A
-----------------------------
1-208.x;1-209.4;;
```

3.3 Datenbereinigung

3.3.1 KOMBI-Kode

Aus OPS Versionen 2.1 und 2.0 wird die erste Zeile der Kode-Datei entfernt, welche den KOMBI-Eintrag enthält.

3.3.2 None statt UNDEF

Für OPS Versionen 2009 bis 2004 wird der Kode-Wert `None` durch `UNDEF` ersetzt, sowohl in den Kodes-, als auch in den Umsteiger-Dateien.

3.3.3 Kreuz-Stern-System

Für die ICD-10-GM Versionen 2.0 und 1.3 werden die Zeichen `+`, `*` und `!` aus den Kode-Werten entfernt – sowohl in der Kodes-, als auch der Umsteiger-Datei.

3.3.4 Punkt-Strich-Notation

Für die ICD-10-GM Versionen 2004, 2.0 und 1.3 wird zuerst die Zeichenfolge `.-` aus den Kode-Werten in beiden Dateitypen entfernt. Danach wird nochmals `-` entfernt. Die Reihenfolge ist wichtig, weil in Version 2004 zum Beispiel Kodes `A00.-` und `G82.1-` vorkommen.

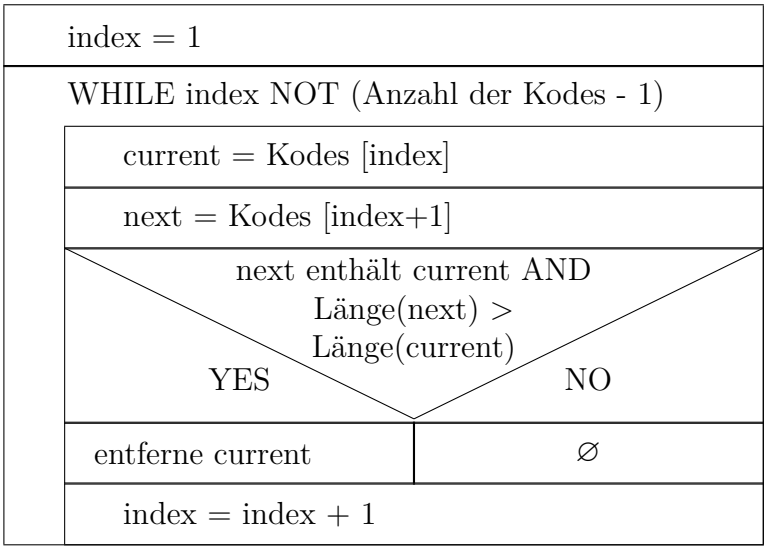
3.4 Instanzreduktion

3.4.1 Umsteiger

Für viele Versionen sind über 90% der Umsteiger-Einträge automatische Überleitungen in den gleichen Kode. Diese müssen also gar nicht in eine Applikation aufgenommen werden, unter der Annahme, dass nicht vorhandene Umsteiger gleich automatische Überleitungen sind. In dem Fall können alle Umsteiger ausgeschlossen werden, bei denen der neue Kode gleich dem alten Kode ist und die automatische Überleitbarkeit in beide Richtungen gegeben ist.

3.4.2 Nicht-endständige Kodes

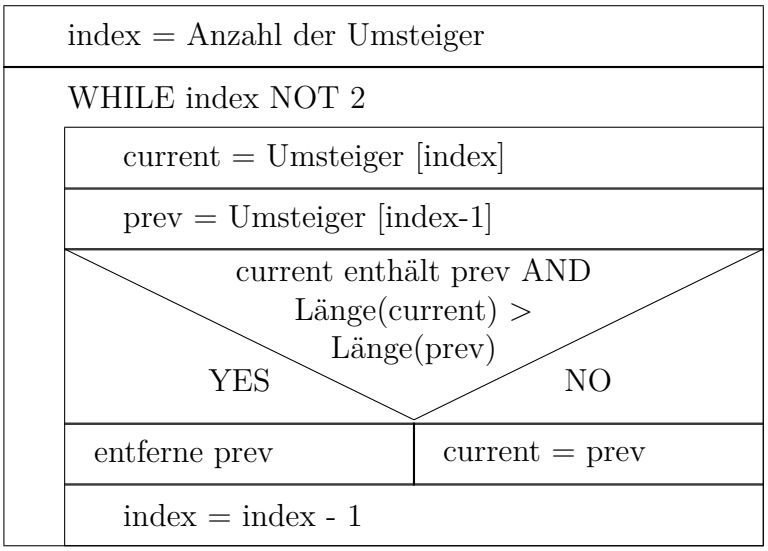
Für die meisten Anwendungen sind eigentlich nur die endständigen Kodes relevant. Statt diese bei jeder Operation herauszufiltern, können beim einmaligen Einlesen der Daten auch einfach die nicht-endständigen Kodes ausgeschlossen werden. Da die Dateien alphabetisch nach dem Kode sortiert sind, geht das mit folgendem Algorithmus[1]:



3.4.3 Nicht-endständige Umsteiger

Die Überleitung von ICD-10-GM Version 2.0 auf 1.3 ist der einzige Fall, in dem die Umsteiger-Datei nicht-endständige Kodes enthält. Durch das Entfernen der Sonderzeichen von Punkt-Strich-Notation und Kreuz-Stern-System gibt es außerdem doppelte Einträge in der ersten Spalte, das heißt bei den Kodes der Vorgänger-Version. Folgender Algorithmus entfernt die überflüssigen Einträge:

[1] Der Pseudocode ist in Struktogrammen nach (Nassi and Shneiderman, 1973) beschrieben. Die später vorkommenden Algorithmen zur Umsteiger-Suche enthalten viele Variablenzuweisungen, die abhängig von einer Bedingung sind und deren parallele Darstellung in Nassi-Shneidermann-Diagrammen übersichtlicher ist. Sie werden also im Sinne der Einheitlichkeit ebenfalls für die einfachen Algorithmen in diesem Abschnitt verwendet.



3.5 Integration zusätzlicher Informationen

Im nächsten Abschnitt wird erklärt, wie ermittelt wird, ob es zu einem Kode einer bestimmten Version Umsteiger in einer älteren oder neueren Version gibt. Um diese zusätzliche Information speichern zu können, werden die Kodes um eine Spalte erweitert.

Literatur

BfArM, *BfArM - Downloads*. [Online]. Available: https://www.bfarm.de/DE/Kodiersysteme/Services/Downloads/_node.html

——, *BfArM - Systematisches Verzeichnis - Kategorie und Kode*. [Online]. Available: https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/ICD/ICD-10-GM/Systematik/kodestruktur.html

——, *BfArM - Systematisches Verzeichnis - Kategorie und Kode*. [Online]. Available: https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/OPS-ICHI/OPS/Systematik/kategorie-und-kode.html

P. Bonnefoy, E. Chaize, R. Mansuy, M. Tazi, and S. Heckel, *The Definitive Guide to Data Integration: Unlock the power of data integration to efficiently manage, transform, and analyze data.* Packt Publishing, 2024.

I. Fernández and J. Manuel, *UTF-8 & Latex Encodings of ISO-8859 (Latin-1) Character Set*, 2022. [Online]. Available: https://www.researchgate.net/publication/359509972_UTF-8_Latex_Encodings_of_ISO-8859_Latin-1_Character_Set/link/6241ab3b5e2f8c7a03452ac9/download

S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *Big data analytics*, vol. 1, pp. 1–22, 2016.

I. Nassi and B. Shneiderman, "Flowchart techniques for structured programming," *ACM SIGPLAN Notices*, vol. 8, pp. 12–26, 08 1973.

D. Thomas and A. Hunt, *The Pragmatic Programmer: Your Journey to Mastery.* Addison Wesley, 2019.