Medical Informatics Initiative

Supporting Project – Central Office of the National Steering Committee

MII Core Data Set

Last revised: March 10, 2017

Draft from the Core Data Set Drafting Group

DISCLAIMER: The official document has been adopted in German language.

Table of Contents

1. Methodology for Creating the Core Data Set

The assorted recommendations for audit queries discussed during the Interoperability Working Group meeting on January 9, 2017 established the point of departure for the creation of the core data set. As part of this process, the data types for modules named in the queries were summarized and recommendations by group members on structuring and semantic markup fleshed out.

The following criteria were decisive for inclusion in the core data set:

- Relevance to research and patient care
- Relevance to the use cases of the consortia
- Availability and ability to be accessed at the locations
- Degree of structuring and the availability of terminologies

Furthermore, during the group's meeting, data types were assigned based on availability to either a basic module that will be made available for audit queries from all of the consortia, or to various extensions that can be made available supplementally depending on the use cases of the individual consortia.

The Core Data Set Drafting Group expanded on this recommendation over the course of several telephone conferences with regard to the following aspects:

- Specification of content
- Reason for inclusion in the basic module or an extension
- Recommendations for structuring and coding
- Recommendations on how to proceed

Due to the limited amount of time, matters needed to be prioritized and individual modules put aside to be dealt with at a later date. All in all, the core data set should be viewed as a work in progress that will need to be completed over the course of the Medical Informatics Initiative and adapted to the current state of development. This also applies to the replacement of national terminologies (e.g., immediately available code lists from billing documentation compliant with §301 of Volume 5 of the German Social Security Statute Book) currently chosen for pragmatic reasons with international terminologies, once their availability for the consortia has been clarified (e.g., in the case of SNOMED CT) and resources are available to establish corresponding mapping.

The refinement of the core data set should take place in close coordination with the Roadmap from the National Steering Committee and the key issues paper "Minimum Requirements for Interoperability" from the Interoperability Working Group. In the case of several modules of the core data set, the section "How to proceed" contains recommendations for continued development in connection with preparatory or auxiliary projects within the context of the Medical Informatics Initiative.

Tips for interpreting this document

- The text makes reference to a "Sec21-data set" several times (data set in conformity with §21 of Germany's Hospital Fees Act, KHEntgG). This is meant as an allusion to the composition of the content and the syntactic structure, not the source of the data. So, for example, a diagnosis data set structured "analogously to §21" can (and should) by all means be used to provide additional outpatient diagnoses or additional diagnoses not relevant for billing purposes.
- The fact that a data type or data element is named in a module of the core data set does not imply anything about whether data from the locations can be made available or turned over in response to a query. What determines this are the data privacy and protection guidelines and the decision-making processes within the consortia and individual locations (e.g. by Use & Access Committees).
- SNOMED CT is referred to in this document several times as a suitable nomenclature for the coding of various data elements. However, the use of SNOMED CT is subject to corresponding licensing, which means that suitable alternatives may need to be sought in this regard.
- In order to identify actions needed for further work, each data type was categorized in the section "How to proceed" according to the following ranked categories:
    1. Data elements already exist in a structured form, and suitable guidelines for their storage and semantic coding have already been established (e.g., elements analogous to a Sec21-data set)
    2. Structured data elements already exist and there are suitable guidelines for their storage and coding, which however have not yet been implemented (e.g., LOINC coding of laboratory analyses)
    3. Data elements exist, but further work on their structure and processing is still required (e.g., medication)
    4. Data elements are relevant, but their availability still needs to be improved and further processing needs to be handled on this basis (e.g., structured attributes from pathology and radiology results)

2. Overview of the Modules and Data Types of the Core Data Set

Modules	Data Types
Basic module	Person
	Demographics
	Case data
	Diagnoses
	Procedures
	Lab results
	Medication
Oncology extension	Tumor data as per the ADT-GEKID*
Diagnostics extension	Pathology results
	Imaging findings
Critical Care Medicine extension	PDMS** data and high-resolution biosignal data
OMICS extension	Genetic tests and sequencing
Biobank extension	Biomaterial data
Structural Data extension	Structural data
Fees and Cost Data extension	Fees
	Cost data (calculation for the InEK)***

* Data set of the Working Group of German Tumor Centers and the Association of Population-Based Cancer Registries in Germany
** Patient Data Management System
*** Institution for the Hospital Remuneration System

3. Data Types on Hold

Modules	Data Types	Status
Diagnostics extension	Microbiology results	On hold
	(Radiological) image data	On hold
OMICS extension	Miscellaneous omics	On hold
Clinical Documentation extension	Vital signs, symptoms, anamnesis	On hold
Studies extension	Study data	On hold
Consent extension	Informed consent information sheets (see results of the Metadata RG and Consent Working Group)	On hold

4. Core Data Set Structure

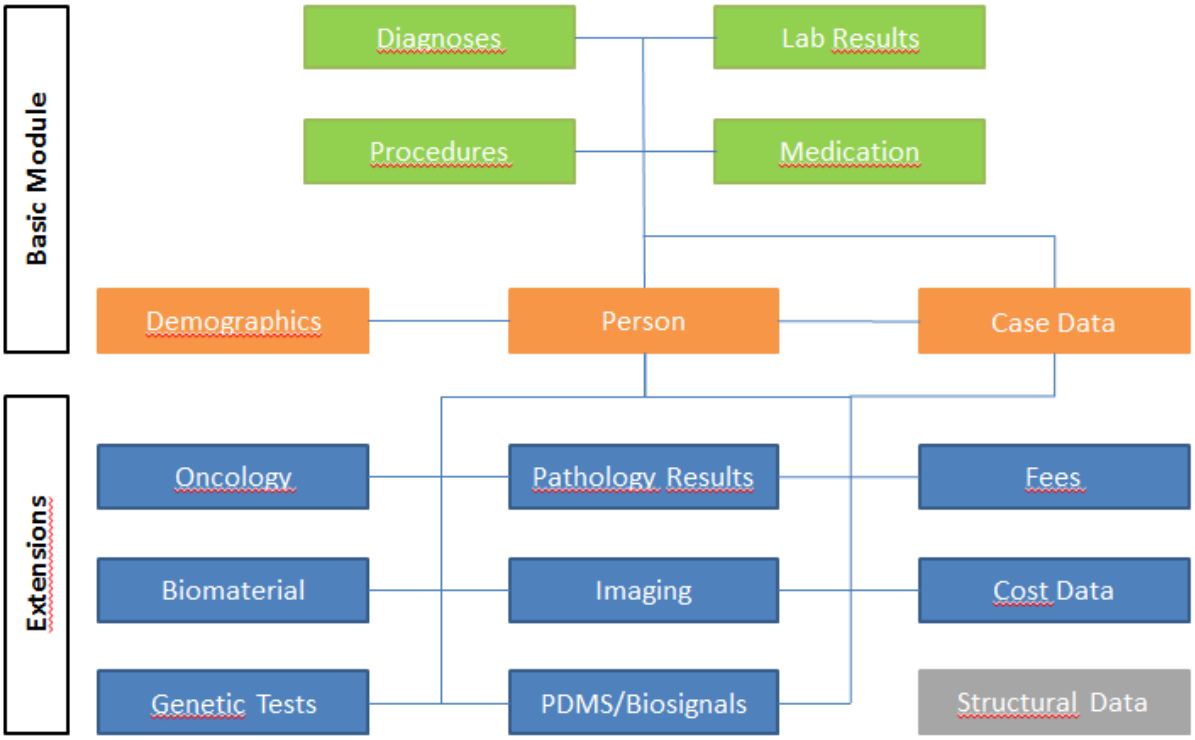

Illustration 1: Block diagram of the core data set

References both to individual treatment cases (and thus also indirectly to the person) and also solely to the person can be produced from the individual modules. Cross-references between the modules (e.g., linking a diagnosis with a medication as an indication) are possible (and desired) but are not yet a component of the core data set. The “Structural data” module contains data that do not refer to a patient and therefore it has no connections to the other modules.

4.1. Basic module: Person

Module	Data Type
Basic module	**Person**

Specification of content

The core data set is to contain a master table entitled PERSON, which will be used to create the references to the other modules. In addition to the use of the locations' internal identifiers, the "Person" table can also contain, as needed, (direct or indirect) personal attributes for (horizontal) integration across multiple institutions and sectors.

Reason for assignment to the core data set

The aim is to develop the core data set such that it will be possible not only to process and integrate the basic data of an individual hospital visit (vertical integration) but also to track the progress of treatment over several visits (horizontal integration). Furthermore, the "Person" module will be able to be used, as needed, for data protection-compliant integration across multiple institutions and sectors.

Recommendations for structuring and coding

In the basic configuration, the "Person" module can contain identifiers from the locations' local identity management systems or unique, pre-existing keys from the other modules of the core data set.
Appropriate data protection measures (e.g., pseudonymization) are to be implemented in connection with any turnover of such.
Over the course of the MII, data protection-compliant approaches are to be developed for horizontal integration, which will then also be able to be implemented in the "Person" module as needed.

Recommendation for how to proceed

Over the course of the project, an internal location identity management system will be put in place at all participating hospitals. Potential implementations of an intra-consortia identity management system will first have to be independently reviewed and developed by each MII partner. A data protection-compliant consortia-spanning identity management system is to be developed in the "TTPs" section of the Roadmap. Based on the information currently available, harmonization of heterogeneous German data privacy and protection law will need to be proposed in several German states.

Stage: 3

4.2. Basic module: Demographics

Module	Data Type
Basic module	**Demographics**

Specification of content

Primary demographic parameters are (date of birth), age, sex, and vital status (living/deceased). Place of residence may also be included in the narrower basic data (as applicable, simplified using the official German Municipality Code ("AGS").

An extended basic data set could include socio-economic attributes, environmental factors, and familial relations.

Reason for assignment to the core data set

The narrower basic data of age and sex are required as minimum information for any age- or sex-related adjustment of reference numbers like risk indicators and survival rates.

Regional codes for place of residence and location of service are required in patient care monitoring for the analysis of regional consistencies and differences.

Recommendations for structuring and coding

Over the course of a lifetime, demographic data may be changeable or unchangeable. With structuring and coding for hospital visits, the narrower core demographic data (minimum basic data set) can be copied into the MII core data set one to one from the "Case" table of the basic data set compiled in accordance with §21 of the Hospital Fees Act (abbreviated: Sec21).

The structuring and coding of core demographic data from the outpatient services sector (potentially also nursing care, emergency services, relation to industry liability insurance/workers' comp, rehabilitation) can be adaptively transformed.

The possibility of adding a separate table for unchangeable demographic data in an expanded core data set should not be ruled out.

Recommendation for how to proceed

In the DIC starting phase, a Sec21-analogous "Case" table holds the key demographic attributes of age, sex, place of residence (ZIP code or "AGS"), and vital status (deceased yes/no). Coordinated adjustments can be made as needed.

Stage: 1

4.3. Basic module: Case data

Module	Data Type
Basic module	**Case data**

Specification of content

Normally, medical documentation differentiates between person - patient - case (Visit_occurrence); additionally there is also a differentiation between administrative and medical cases.

For the basic module of the MII core data set, population of a "Case" table analogous to the Sec21-data set and the "Visit_occurrence" table in the OMOP common data model is a good option. For the billing of inpatient cases, the Sec21-data set is based on the administrative definitions. A "case" begins with admission into the hospital on an admission date and ends with the discharge on a discharge date. In this regard, case data include various 1:1 attributes of a visit, among others the type of discharge with the possible attribute value "expired". Valuable demographic attributes in the "Case" table are age, sex, and place of residence (ZIP code). The insurance policy number and primary service provider are also included. During a stay in the hospital, one department is primarily responsible for each case at all times. This is rather precisely documented by the "DEPT" table in the data set compiled in accordance with §21 of the Hospital Fees Act, and can be copied on this basis into the MII core data set to supplement the case data. In the outpatient services sector, depending on the type of fee, a person's cases are regularly separated from one another by the start and end dates of calendar quarters. Nevertheless, the greatest possible mapping similarity should be aimed for in the core data set.
For quarterly cases, the actual physician consultations could comprise a corresponding supplement to the case data for the outpatient services sector.

Reason for assignment to the core data set

The case data (attributes in the "Case" table) are required in particular for assignment to a sector, date(s), and location and for the provenance of the data. They form the backbone of the data model.

Recommendations for structuring and coding

The structuring and coding of the "Case" table may lean heavily on the "CASE" and "DEPT" tables of the Sec21-data set.

For outpatient cases, efforts should be made toward an integration in this format taking into account a special provenance attribute.

Recommendation for how to proceed

In the DIC starting phase, Sec21-analogous "CASE" and "DEPT" tables hold the most important case data. Coordinated adjustments can be made as needed.

Stage: 1

4.4. Basic module: Diagnoses

Module	Data Type
Basic module	**Diagnoses**

Specification of content

In the health care system, diagnoses constitute the reason for treatment and are a key categorization attribute.

In the inpatient services sector, the primary and additional diagnoses are compiled in the hospital information systems (HIS) for multiple purposes, including for export in the form of basic data compiled as per a Sec21-data set.

In the outpatient services sector, each treating physician normally documents only one "quarterly diagnosis" in the billing data (= existing routine data) per case.

Reason for assignment to the core data set

The diagnoses are (usually) the most important independent variable for many issues.

Recommendations for structuring and coding

In the MII core data set, the structuring and coding of a "Diagnoses" table as a sub-table of the "Case" table (Visit_occurrence) can for the most part be based on the "ICD" table in the inpatient Sec21-data set.

The aim should be to add the "present-at-admission" and "present-at-discharge" attributes or a period of validity (timing-setting) for each diagnosis. The exclusion – pursuant to the German Coding Guidelines – of additional diagnoses in the Sec21-data set that are fee-free and do not conform to billing requirements should not be used in the MII data set (neutralization). Coding in accordance with SNOMET CT should supplement the ICD coding (internationalization) in short order.

For the inclusion of diagnosis information from the outpatient services sectors, efforts should be made to use a format similar to that for inpatient services sector.

Recommendation for how to proceed

A "Diagnoses" table ("Condition_occurrence") analogous to the "ICD" table of the Sec21-data set should be planned in the DIC starting phase.

Efforts should be made toward speedy adjustments for setting of timing, neutralization, and internationalization. Further coordinated adjustments will be made as needed.

Stage: 1

4.5. Basic module: Procedures

Module	Data Type
Basic module	**Procedures**

Specification of content

Procedures constitute major service bundles of diagnostics and treatment. The most significant procedures are operations. For fee-related reasons, a few expensive medications are also listed in the inpatient Sec21-data set.

Whether to allocate medications to "Procedures" or to maintain them in a separate "Medications" table will have to be decided over the course of development. A separate decision will also have to be made regarding the extent to which the "Procedures" table will also be maintained for outpatient cases.

Reason for assignment to the core data set

In the health care system, procedures comprise the most significant interventions and are among the decisive influential factors for many issues.

Recommendations for structuring and coding

The structuring and coding of a "Procedures" table can for the most part be based on the "OPS" (German procedure classification) table of the Sec21-data set. Efforts should be made toward transformation of service information for outpatient cases into a "Procedures" table.

Recommendation for how to proceed

A "Procedures" table ("Procedure_occurrence") analogous to the "OPS" table of the Sec21-data set should be planned in the DIC starting phase. Coordinated adjustments can be made as needed. Overlaps and the differentiation of procedures and medications should be reviewed in short order.

Efforts should be made toward internationalizing and homogenizing outpatient and inpatient service documentation.

The integration of service information from supplemental sources such as tumor documentation (there: operations, radiotherapy, chemotherapy, targeted therapy, immunotherapy) would be goal-oriented. Any diagnosis-therapy-relations there should be preserved (copied).

Stage: 1

4.6. Basic module: Lab results

Module	Data Type
Basic module	**Lab results**

Specification of content

Laboratory tests are performed for almost every patient receiving inpatient care; most of the resulting lab result data are available centrally. For each lab test, the following information should be transferred to the DIC with reference to patient and case:

- Test performed – name of analysis with unique test ID (using LOINC, see below)
- Test date
- Test result (measurement) – with standardized unit (using UCUM, see below)
- Interpretation: Identification, whether pathological value (recommended)
- Scale type (optional)
- Reference range (recommended)
- Originating laboratory (optional)

Outside of the range of classic clinical-chemical and hematological lab data, other clinical tests, microbiological findings, and vital signs can be mapped in the same way. LOINC standardization and standardized data use can gradually be extended to other observations (e.g., in radiology, pathology) through to descriptive medical documentation (LOINC Clinical Document Ontology).

Reason for assignment to the core data set

To a large extent and for a very large percentage of cases, this body of data can supplement the databases in conformity with §21 of the Hospital Fees Act and help to answer medically relevant questions (e.g., related to pharmacovigilance, symptomatic screening, evaluation of medical documentation/diagnosis, treatment monitoring, infection research, clinical study inclusion and exclusion criteria, development and validation of new reference ranges). Early inclusion and early use of this body of data appear promising and practical, since at least in the first step, the standardization (subset clinical-chemical lab test, see below) does not involve manually performed documentation processes but instead can be computer generated. Furthermore, the DIMDI (German Institute of Medical Documentation and Information) has provided a German translation for a majority of the LOINC.

Recommendations for structuring and coding

LOINC + UCUM (laboratory parameters and units)

a) LOINC:
LOINC (Logical Observation Identifiers Names and Codes) comprises a flat table with unique international ID codes for clinical tests and observations that allow for a high degree of discrimination between test

variants. LOINC is freely available and undergoes maintenance regularly (Regenstrief Institute, U.S.A.). For details, see the LOINC website [1] (together with database and tools). LOINC is explicitly mentioned as reference terminology in the FHIR resource "Observation", among others.

Consequently, LOINC standardization of tests obtained with computer assistance is especially simple since the master table of analyses/tests in the associated IT system (e.g., LIS) only has to be expanded one time by one column for the associated LOINC IDs – and these can be provided together with the analyses via an interface. Since the transmission of multiple IDs already complies with the HL7v2 protocol, there is a high probability here of realization using legacy software.

b) Selection of a subset with LOINC standardization
The LOINC subset for the first stage of standardization for all of the DICs and consortia will follow on the heels of these previous experiences:

- the IHE LOINC Test Codes Subset (from IHE LAB-TF4)[2]: 2,500 codes
- the previous experiences of German university hospitals (especially the University Medical Center Schleswig-Holstein in Kiel since 2001)
- the determinations of the national Electronic Health Record project "ELGA" in Austria: 3,190 codes[3],[4]
- KAV (Vienna Hospital Association) LOINC "gold list": 2,065 Codes[5]
- LOINC Top 2000+ Lab Observations[6]
- LOINC Top 300 Orders[7]
- LOINC Order Code S&I Framework Initiative ("aLOINC"): 1,529 Codes[8]
- LOINC Codes for Common CDISC tests: 500 Codes[9]

Efforts should be made to establish about 1,000 parameters (comparable with the quantity set in ELGA) as a practical subset size that allows for ample use (not just for the 2020/21 audit). However, this number is not a milestone per se, rather the feasibilities and circumstances may need to be adjusted (in one or the other direction).

For guidance: The IHE LOINC Test Codes Subset comprises 2,500 codes broken down as follows:

- Discipline - Number of LOINC test codes selected
- Chemistry including urinalysis and challenge studies - 873
- Hematology - 284
- Toxicology + drug monitoring - 194
- Virology (including serology) - 374

[1] www.loinc.org
[2] https://www.ihe.net/Technical_Framework/upload/ihe_lab_TF_rel2_1-Vol-4_FT_2008-08-08.pdf
[3] http://www.elga.gv.at/technischer-hintergrund/technische-elga-leitfaeden/index.html
[4] http://www.hl7.at/wp-content/uploads/2013/10/ELGA-Value-Sets_3.4.xlsm
[5] http://www.bmgf.gv.at/cms/home/attachments/1/3/0/CH1064/CMS1240821423857/dokuambulant_handbuch_laborleistungen_kal_2011_gesamt.pdf
[6] http://loinc.org/downloads/usage/obs
[7] https://loinc.org/usage/orders
[8] http://wiki.siframework.org/file/view/aLOINC+Order+Code+final+report+3+31+2015.docx
[9] http://loinc.org/discussion-documents/CDISCcommonLOINCtests20050214.pdf

- Parasitology and mycology - 158
- Bacteriology - 387
- Immunology and cell mark - 278
- Patient and specimen - 30

A subset of this size allows for coding of about 99% of all lab tests in normal operations.

c) UCUM
The UCUM (Unified Code for Units of Measure) defines a uniform, international machine-readable reproduction of units of measurement.[10]

The implementation of UCUM in legacy LIS systems is somewhat more complex. In this regard, an exchange among experts (workshop) is recommended specifically to discuss experiences in neighboring countries (in particular Switzerland, Austria, the Netherlands, and the U.S. if applicable). At present, there is no known routine implementation of UCUM in Germany. Of course, mapping to UCUM units can also take place at the level of a research database.

d) Syntactic standardization of communications of lab test and other investigation data
To the greatest extent possible, the patient- and case-related communication of lab test and other investigation data using LOINC (and, as applicable, UCUM) standardization is to be independent of the choice of syntax; nearly all relevant communications standards and profiles support the use of LOINC (and UCUM):

- HL7 CDA
- HL7 v2 Messaging
- HL7 FHIR
- IHE
- LDT 2.0 and 3.0
- CDISC SDTM

Use of international standards is preferable at the hospital, but harmonizing the ETL process for the import of lab data into the DICs is not necessarily required across all of the consortia.

Recommendation for how to proceed

a) LOINC subset:
In the first step, a **subset** is to be selected guided by the following criteria:

- The test (analysis) exists at the highest possible number of locations participating in the MII
- The test is relevant for the highest possible number of patient cases
- The test is relevant for research and patient care issues
- The test is relevant for the use cases of the consortia
- LOINC codes exist (global tests ahead of special tests)
- Easy handling of LOINC standardization for the relevant test type
- LOINC code already exists as an approved part of other germane projects
- (LOINC name is available in German)

[10] http://unitsofmeasure.org

Based on the criteria, it would be logical to start with a **subset of clinical-chemical lab tests (including basic hematology/toxicology)** and gradually expand this in connection with the Roadmap to include more complex areas (e.g., microbiology). The subset from the first stage should be available, standardized, at all DICs by **January 1, 2020** for every included patient for whom analyses from the subset have been performed.

- Development of a recommendation for a German LOINC subset with special consideration given to previously existing work (II/2017)
- If needed: **Possibly an experts' workshop** with the authors of analogous LOINC subsets in Austria (ELGA), Switzerland (e-health Suisse), France (IHE LTF-4), possibly the Netherlands and the U.S.
- A small drafting group is to send a **subset draft for the German "LOINC 1000+"** codes to be coordinated to all locations filing requests so that they can take these under advisement and have their IT and clinical pathology departments comment.
- **Final determination as early as III+IV/2017**, so that two years are available for mapping and for setup at the locations.
- **Mapping** of the "LOINC 1000+" at all DIC locations between January 1, 2018 and December 31, 2019
- The subset from the first stage is to be available, LOINC-standardized, at all DICs by **January 1, 2020** for every included patient for whom analyses from the subset have been performed.
- The body of data can be used for generally accessible audit queries starting IV/2020.

Stage: 2

b) UCUM:

- **Experts' exchange (workshop)** specifically to discuss experiences in neighboring countries (in particular Switzerland, Austria, the Netherlands, the U.S. if applicable) with the implementation and use of UCUM in legacy systems. Important: Industry needs to be included (especially manufacturers/providers of LIS, PDMS, and "KAS" (Clinical Positions System))
  Target period: III+IV/2017
- Based on this, further procedural recommendation regarding supplementation of the Roadmap. (to be completed by the start of the setup phase)

Stage: 3

4.7. Basic module: Medication

Module	Data Type
Basic module	**Medication**

Specification of content

The types of documentation of the dispensing of drugs can be categorized as follows:

1. Medication in the hospital (primarily inpatient/partial hospitalization)
2. Discharge medication
3. Outpatient medication
4. Self-medication (OTC)
5. Medication associated with clinical studies

Information on medication can range from simple documentation of the dispensing of a drug in a treatment case, to a detailed, structured record of individual doses with coding of active ingredient, form of administration, route of administration, and dosage in accordance with established international standards.
A minimum form for type 1 can be achieved by all inpatient care facilities based on OPS codes for medications eligible for supplemental reimbursement. Furthermore, in intensive care units, completely structured medication documentation is normally completed in a PDMS, which is sometimes also used in normal inpatient care in connection with visit documentation systems or dedicated prescription systems. Moreover, case-related documentation is frequently completed in hospital pharmacy systems, e.g., in connection with internal preparation of intravenous solutions or batch documentation.
In future, data for types 2 and 3 can be made available via data from the Medication Plan[11]. Starting in 2018, medications are to be retrievable via the electronic medical record. For type 4, no patient-related documentation is currently foreseen (in the case of patient-administered, also included in the Medication Plan; for the long-term conceivably via patient portal). Study medications (type 5) are frequently structured in electronic data capture systems but recorded without semantic storage (save for the coding of side effects in MedDRA as mandatory components of the pharmacovigilance reporting chain). Limitations may arise here due to the blinding of study medications.

The active ingredient should be the minimum retrievable scope of information for a medication.

Furthermore, the following data elements are to be made available in a further expansion stage:

- The drugs' trade names
- Dosage with unit of volume
- Form of administration
- Site and route of administration

11 https://www.abda.de/fileadmin/assets/Medikationsmanagement/BMP_Anlage3_Spezifikation_Version_2_3_final.pdf

Where corresponding data is available, the following data elements can also be added in a further expansion stage:

- Dosing schedule
- Composition and rate of administration of intravenous therapies and infusion pumps
- Indication (in terms of the reference to an original diagnosis)

Reason for assignment to the core data set

The prescription of medications is a core process of routine care and takes place at all hospitals participating in the MII. Nevertheless, the percentage of digitally documented prescriptions differs widely across the locations with regard to the degree of structuring as well as the populations and medications covered. At this time, there is no general overview of the availability of structured medication data at the consortia locations. This fact is taken into account by way of the recommended multi-stage process to achieve availability and the recommended preparatory project.

Medication data are of key significance for a number of issues, e.g., in pharmacovigilance and as inclusion and exclusion criteria for study populations. In the audit queries put together by the Data Sharing Working Group, they will be listed in the recommendations from the consortia HD4CR and SMITH.

Recommendations for structuring and coding

Internal hospital medication (type 1) can be made available as a "minimum form" for medications eligible for supplemental reimbursement uniformly across the country on the basis of the OPS codes. Active ingredients should be coded in accordance with the WHO's freely available ATC classification system (available from the DIMDI)[12]; furthermore, specific drugs can be referenced by their German pharmaceutical barcode ("PZN") or their trade name. The Federal Institute for Drugs and Medical Devices (BfArM) maintains the freely available Catalogue of Pharmaceutical Ingredients ("ASK")[13], which has been supplemented with an active ingredient designation in view of possible use in the Medication Plan[14]. The UCUM[15] is to be used where practical for coding units of volume for dosage. Value sets from SNOMED CT are available (subject to licensing) for sites and routes of administration. Medications are also included in LOINC and MeSH, however there are no significant databases in Germany known to be coded using them. Furthermore, for the structuring of medication data, reference is made to the HL7 FHIR resource "MedicationAdministration"[16].

Recommendation for how to proceed

In an initial expansion stage, data on medications eligible for supplemental reimbursement (as per OPS class 6, type 1) should be available by the end of 2018. By 2020, detailed structured data (for subpopulations such as, for example, intensive care units, as applicable) and on known medications should be made available in connection with the Medication Plan.

[12] http://www.dimdi.de/static/de/klassi/atcddd/
[13] http://abdata.de/datenangebot/abda-datenbank/pharmazeutische-stoffliste/
[14] http://www.pharmazie.com/dacon32/Support/ABDATA_Stoffkatalog_2014.pdf
[15] http://unitsofmeasure.org/trac
[16] https://www.hl7.org/fhir/medicationadministration.html

To reach the second expansion stage, a preparatory project with the objective of determining the actual status of the medication data locally available at the locations and of specifying the details of the data structures is recommended.

With regard to the fundamental importance of medication safety for patients, the topic of medication is also suited to consortia-spanning use cases in connection with an auxiliary project.

Stage: 3

4.8. Oncology extension: Tumor data in accordance with “ADT-GEKID”

Module	Data Type
Oncology extension	**Tumor data as per the “ADT-GEKID”**

Specification of content

Tumor data cover the following areas, among others:

- Characterization of the tumor disease with information on the organ, local dissemination, and the lymphatic and hematogenous metastasis
- Information on the course of treatment (surgical, systemic, radiotherapy)
- Information on the continued course of the disease and on survival

The content of the “Oncology” extension will overlap with the content of the “Demographics”, “Diagnostics”, “Procedures”, and “Medication” modules. It nevertheless offers a significant added value due to the tumor-related compilation of various items of diagnostic and therapeutic information. These cannot be readily presented by the other modules together with the tumor aspects. Irrespective of this, an additional transfer of data from the “Oncology” module to the respective basic data types may be practical in order to achieve complete mapping, e.g., of all procedures (including those otherwise potentially documented only in an oncological context).

Reason for assignment to the core data set

Tumor diseases play a key role in university medicine and are recorded in a generally accessible structured form and followed up on in line with all existing cancer registry laws. Corresponding documentation is available at all hospitals with certified oncological centers in quality-assured form.

Recommendations for structuring and coding

The Working Group of German Tumor Centers (ADT) and the Association of Population-Based Cancer Registries in Germany (GEKID) have developed a joint "ADT-GEKID" basic data set [17] that allows for a structured description of tumor diseases. Furthermore, organ-specific extensions are available or in the process of being developed.

Recommendation for how to proceed

Integration of the raw data available through ADT-GEKID in the DICs preserving the tumor reference of the data elements (additional categorizations by person and case as applicable) Stage: 1

4.9. Diagnostics extension: Pathology results

Module	Data Type
Diagnostics extension	**Pathology results**

Specification of content

Pathology results are an essential medical record component for oncology in particular, but of course not just for this field. Because the structuring and content of pathology results are very heterogeneous depending on the organ under investigation, initially limiting pathology results to oncological use cases is recommended. As a rule, the results are stated as free text. If structured information is extracted via text mining, the origin needs to be made clear for purposes of provenance.

Reason for assignment to the core data set

Oncological use cases are addressed in several consortia; in these the pathological results are an essential source of information. As for content, there are some overlaps with the tumor data mapped in the ADT-GEKID.

Recommendations for structuring and coding

Due to what has generally been an absence of structuring up to this point, free texts are initially to be allowed with structured metadata (IHE-XDS). In a further expansion stage, efforts should be made toward structuring based on HL7 PaLM[18]/IHE APSR[19] including the value sets described there, which are already mapped with PathLex, even as the report is being generated.

[17] http://www.tumorzentren.de/onkol-basisdatensatz.html

[18] https://art-decor.org/art-decor/decor-project--psr-

[19] http://www.ihe.net/Technical_Framework/upload/IHE_PAT_Suppl_APSR_Rev1-1_TI_2011_03_31.pdf

How to proceed

Expansion stage 1: Definition of specific metadata for the free texts Expansion stage 2: Definition and development of a subset from HL7 PaLM/IHE APSR for structured and coded information in the context of tumor diseases together with the consortia reporting corresponding use cases Stage: 4

4.10. Diagnostics extension: Imaging findings

Module	Data Type
Diagnostics extension	**Imaging findings**

Specification of content

What are relevant here are structured reports of all imaging procedures where report templates are used and the report is stated as free text. The templates are currently undergoing the approval process for various issues (e.g., rectal carcinoma, colonic carcinoma, solid pancreatic carcinoma, etc.) within the German Radiological Society (DRG) and will be made available to the public.

Reason for assignment to the core data set

Imagining is one of the most important diagnostic procedures for all fields of medicine. The structured recording of findings leads to fully answering all questions of relevance to the issue. There will be criteria set (in consensus with the referring physicians) for the evaluation. Ideally, a dictionary with clear definitions for the vocabulary to be used will be available to the various entities. Good examples of this are BI-RADS for diagnostic breast imaging or PI-RADS for diagnosis of prostatic carcinoma. The structured recording of findings results in better, quantified recording and traceability of the individual findings acquired.

Recommendations for structuring and coding

FHIR DiagnosticReport[20] and FHIR ImagingStudy[21] (DICOM Integration); terminologies to be used in this connection are LOINC and RadLex; RadLex is the standard for radiological terminology and is currently

[20] https://www.hl7.org/fhir/diagnosticreport.html
[21] https://www.hl7.org/fhir/imagingstudy.html

being translated into German by the German Radiological Society (will be available in German starting in Q1/2018). Analogous to RadReport.org[22] from the RSNA, a home page for the templates that will have relevant content will be hosted by the German Radiological Society no later than by Q1/2018 (report templates for all significant tumor entities). Beyond the report templates, DICOM SR offers the opportunity to work using hyperlinks between (radiological) imaging material and the report text.

Recommendation for how to proceed

Expansion stage 1: Free texts made available
Expansion stage 2: Access provided to selected structured attributes
Stage: 4

4.11. Critical Care Medicine extension: PDMS data and high-resolution biosignal data

Module	Data Type
Critical Care Medicine extension	**PDMS data and high-resolution biosignal data**

Specification of content

PDMS data are comprised of attribute categories such as

- Structural data on the facility, including bed count, number of employees (including qualifications), record of quality indicators, … (according to the core data set of the DIVI (German Interdisciplinary Association for Critical Care and Emergency Medicine))
  => see Structural Data **extension**
- General information such as age, sex, diagnoses, procedures, length of stay, referring discipline, days and location of stay in the hospital before admission to an intensive care unit, ...
- Physiological data such as systolic blood pressure, heart rate, respiratory rate, body temperature
- Scores such as the Glasgow Coma Scale, SAPS II, SAPS III, SOFA, CAM-ICU, …
- Mechanical ventilation values such as ventilation mode, plateau airway pressure, FiO2, oxygenation (PaO2/FiO2), ...
- Blood test results from point-of-care devices such as blood gas analysis (pH, paO2, ...), blood sugar, electrolytes, lactate, …
- Data from laboratory medicine, e.g., blood count, coagulation values, creatinine, bilirubin, ...
  => see **Basic module:** Lab results
- Blood products such as erythrocyte and thrombocyte concentrates (number of units)
- Medications such as dopamine, adrenaline, etc., including type of administration (bolus,

[22] http://www.radreport.org/

intravenous infusion pump, gaseous form, …), dosage, time, …
- Fluid balances as they relate to excretions of urine, blood, ... (ml)
- Monitoring data and vital signs such as blood pressure, heart rate, temperature, central venous and pulmonary arterial pressure, ...
- Structured information from shift reports and/or follow-ups by means of text mining, e.g., regarding diagnoses made, complications, ...
- High-resolution biosignals (to be clarified: which device data in what resolution are of interest over what period of time)
- Structured information from electronic microbiology reports (antibiograms and resistances) Preferably, this should already be extracted in the HIS area by means of text mining before being transmitted to a PDMS or, respectively, DIC.
  => see Diagnostics **extension**

Reason for assignment to the core data set

Certain PDMS data refer to data in other modules, among others Sec21-data set, lab results, microbiology or radiology results. From the perspective of the MII, their special significance relates to both the severity of the patients' conditions and the capture of granular data in specialized documentation systems.

Because of the high amount of structured data, PDMS data are an ideal source for the German Federal Ministry of Education and Research Medical Informatics challenge "Secondary Use of Health Care Data for Research". Three consortia (ADMIRE, SMITH, share-it!) are to describe a use case in the field of "critical care medicine".

Aggregates (e.g., MIN, MAX, AVG, SUM) of individual parameters relating to particular points in time or intervals, e.g., highest and lowest blood pressure or creatinine value and total bilirubin in a 24-hour period after admission (or the time the diagnosis was made or the patient was transferred to the intensive care unit) are often of interest for the use of PDMS data for secondary uses such as quality assurance or clinical research. This requires that the time be recorded with precision, which means it is also necessary to synchronize the time on devices.

Recommendations for structuring and coding

For the field of critical care medicine, there are a series of standards and initiatives for improved structuring and coding of content, which ideally should already be implemented in the PDMS:
- The "critical care medicine" data set (2010)[23] from the DIVI and DGAI (German Society of Anesthesiology and Intensive Care Medicine); primarily for the evaluation of quality assurance measures and health services research
- SAPS II- and Core-10-TISS scores obtained in connection with complex critical care documentation[24]

[23] http://bit.ly/2l70ZyH

- AKTIN project[25] (national ER admission registry): An HL7 CDA specification[26] was prepared for the "ER admission" core data set from the DIVI used here (containing overlaps with the "critical care medicine" core data set mentioned above), including the use of a number of vocabularies, among others LOINC, UCUM, and in certain cases SNOMED CT. The CDA standard PHMR (Personal Healthcare Monitoring Report) on the communication of device data n[27] is also due a mention here.
- The Ministry of Education and Research's project OR.NET[28] (integration and networking of medical devices with one another and with adjacent IT systems): The ISO 11073 standards family has been further refined, including a data model (among other things, device description with data provenance aspects) and a nomenclature for device data[29]. Waveforms and streaming data for mapping high-resolution data are also addressed as are transformations of the device data pursuant to HL7 FHIR. There are international efforts to smooth these results in the context of IHR. => see mapping "ISO/IEEE 11073 - IHE PCD DEC - HL7 FHIR"[30]
- IHE domains "Patient Care Device (PCD)"[31]: Profiles such as Device Enterprise Communication [DEC] or Rosetta Terminology Mapping [RTM][32] exist for the interoperability challenges of "critical care medicine". The DEC profile (previously device data => HL7 V2 OBX segments) is to be updated with HL7 FHIR in mind (see above). The RTM profile with specific device terminology is to be harmonized with LOINC/UCUM and also SNOMED CT.
- Alternatively, waveforms including the metadata can be communicated in standardized form with DICOM or, respectively, DICOM SR (Structured Reporting)[33].

Recommendation for how to proceed

Stage 1: The DIVI core data set "critical care medicine" is supported by all PDMS manufacturers, i.e., the attributes defined here are available in structured form.

Of course, numerous PDMS attributes such as "body temperature (highest)" or "total bilirubin" are captured in aggregated form within one hour before or after ICU admission. The aim is the level of raw data as a starting point for these types of aggregations and a description of the data elements using structuring and terminology standards. Raw data elements to be made available should also cover the source data of the SAPS and TISS scores.

Stage 2: Due to the large number of possible parameters and attribute values as well as the tight intertwining with other data types, agreement on a universal attribute catalogue for PDMS at the level of individual data elements is difficult. It would appear to be practical to coordinate a subset of relevant data elements within the three consortia with critical care medicine use cases **over the course of 2017**. Keeping

[24] https://www.dimdi.de/static/de/klassi/faq/ops/kapitel_8/ops-anleitung-intensivmedizin-8009.pdf
[25] http://www.aktin.org/
[26] http://wiki.hl7.de/index.php?title=IG:Notaufnahmeregister
[27] https://www.ncbi.nlm.nih.gov/pubmed/22874265
[28] http://www.ornet.org/
[29] http://bit.ly/2lMj0RJ
[30] http://bit.ly/2l4OLq8
[31] https://www.ihe.net/Patient_Care_Devices/
[32] https://rtmms.nist.gov/
[33] http://dicom.nema.org/dicom/supps/sup30_lb.pdf

in mind the core data sets identified, a determination must be made as to which meta attributes are necessary for each data element, e.g., semantic annotations, data provenance, etc.
Previous work should be consulted for this, e.g.,

- The above-mentioned AKTIN project with HL7 CDA specifications and work toward the mapping of vital signs (e.g., blood gas analysis) using LOINC, UCUM, SNOMED CT[34]
- Existing HL7 FHIR data structures for device data based on the IEEE 11073 data model and nomenclature

Stage 3: How to proceed from there is still unclear for several data types. Starting at the beginning of the setup phase (I+II/2018), separate working groups are to develop solutions for possible problems:

- Existing standards (IEEE 11073) including protocol implementations are not put into practice by manufacturers => proprietary data => primarily a governance problem!
- High-resolution biosignals are rarely captured in PDMSs. Instead, averaged, pre-processed device data or, as the case may be, data contaminated by artifacts are transmitted.
  => coordinated approach, see above-mentioned standards
- Extraction of structured data from electronic texts by means of text mining:
  => cross-PDMS concept, e.g., relating to detection rates, ... (data provenance).

4.12. OMICS extension: Genetic tests/sequencing

Module	Data Type
OMICS extension	Genetic tests/sequencing

Specification of content

Genome-wide analyses allow genome characteristics such as gene variations (alleles) and spontaneous changes in the genome (mutations) to be measured and presented. In the case of variants based on mutations, there is generally a differentiation between mutations in the germline (germline mutations), which are heritable, and those in somatic cells (somatic mutations, e.g., in tumor cells). Meanwhile, there are a number of measuring techniques available: gene panels, exome and whole genome sequencing. The processing of raw data is accomplished via so-called “bioinformatic pipelines”, which are likewise heterogeneous.

Reason for assignment to the core data set

Genetic tests provide information on causal relationships between structural variants or changes in the genome and potential diseases and possible therapies. Thanks to this and to the availability of inexpensive tests, they have become very important to the field of medicine. They are used for diagnostics, differential

[34] http://www.egms.de/static/de/meetings/gmds2008/08gmds203.shtml

diagnostics, and to select therapies. Furthermore, they provide prognostic and predictive information regarding specific therapies (e.g., pharmacogenomics). Meanwhile, genetic tests are important in all disciplines of medicine; they are used ubiquitously both in routine diagnostics (e.g., typing tumors) as well as in research.

Recommendations for structuring and coding

The FHIR profile "Standard Profile for Genetics"[35] should be used for guidance. This profile is an extension of the "Observation" module; building on this, all of the data elements defined there can be accessed (e.g., specimen):

"Observation-genetics-profile (i.e. Standard Profile for Genetics) extends Observation[36] resource to enable reporting of structured genetic test results. In addition, the genetics profile contextualizes well established standards from the field of clinical genetics into the standards of healthcare (e.g. HGNC - HUGO Gene Nomenclature Committee's international standard for gene names, symbols, and identifiers).
Genetic Standards and Resources include:

- Variant Databases: dbSNP, ClinVar, and COSMIC
- Reference Sequences: RefSeq and ENSEMBL
- Gene Symbols and Identifiers: HGNC - Human Gene Nomenclature Committee
- Variant Nomenclature: HGVS nomenclature from the Human Genome Variation Society
- Variant Feature Annotation: Sequence Ontology (SO) and LOINC
- Locus: Gene"

This allows all of the important information on genetic tests to be mapped, and all of the important data elements and terminologies are indicated; however, there is no limitation to specific entities such as gene loci.

Recommendation for how to proceed

[TBA]
Stage:

[35] https://www.hl7.org/fhir/observation-genetics-cg-prf-1a.html
[36] https://www.hl7.org/fhir/observation.html

4.13. Extension: Biomaterial Data

Module	Data Type
Biobank extension	**Biomaterial data**

Specification of content

Biospecimens are to be assembled prospectively in the form of disease-specific and population-based biobanks. Both the upper-level biobank and the individual specimens will need to be described in structured form to be of practical use. Relevant attributes for the collection of specimens include, among others, the symptoms, specimen types, quantity data, and specimen collection sites covered (this in the form of metadata). Specimen-specific data would contain information on specimen type, specimen material, quantity, how the specimen was obtained, pre-analytic processing (including aliquoting, pooling), and storage. Clinical data on the specimen will expressly not be covered by this module but rather should be made available via the modules planned for the respective data type.

Reason for assignment to the core data set

The collection of biomaterials is of ever-increasing importance for medical research. As part of this, the standardized characterization of collections, specimens, how those specimens were obtained as well as their processing is a key prerequisite for successful use. Relevant standards and infrastructures are being refined and disseminated in both national (e.g., German Biobank Node) and international projects (e.g., BBMRI ERIC).

Recommendations for structuring and coding

Additionally, **SPREC** (Standard PREanalytical Code)[37] is available for the characterization of individual specimens and their pre-analytic processing. For each specimen, SPREC describes the specimen type and pre-analytical quality in the following data elements (with standardized value lists):

- type of sample (3 digits)
- fluid biospecimens:
  - type of primary container (3 digits)
  - pre-centrifugation delay and condition (1 digit)
  - centrifugation (1 digit)
  - 2nd centrifugation (1 digit)
  - post-centrifugation delay (1 digit)
- solid biospecimens:
  - warm ischemia time (1 digit)
  - cold ischemia time (1 digit)
  - fixation/stabilization type (3 digits)

[37] http://www.isber.org/?page=SPREC

- fixation time (1 digit)
- (Long-term) storage conditions (1 digit)

At least a selection of these coded specimen descriptions could be used with patient or case reference for the core data set in a standardized form (e.g., type of sample, storage conditions, possibly type of primary container or fixation/stabilization type). To do this, there would first need to be a binding stipulation for all of the biobanks (e.g., via GBN/GBA or the TMF Biomaterial Bank Working Group (AG BMB)) to describe all specimens uniformly using SPREC and to make this information available via interfaces.

MIABIS (Minimum Information About Biobank Data Sharing)[38] provides structures for the characterization of samples collections and studies (MIABIS CORE). More extensive structures, e.g., for the description of individual specimens, are currently being developed within MIABIS. Thus this standard currently (still) falls within the domain of metadata.

Recommendation for how to proceed

Joint planning project with GBN/GBA (and therefore dovetailing with BBMRI) following up on the white paper "Biobanking in the Medical Informatics Initiative" from the GBN/GBA, TMF-AG BMB, and DZG in order to coordinate a usable, binding basic data set for *specimen availability information* and *specimen description data*.

Target time period: The beginning would be at the start of the setup phase (I+II/2018). As applicable, first hold a **joint workshop with all initiatives** (in IV/2017 - e.g., in connection with the National Biobank Symposium - or in I/2018).

Stage: 3

[38] https://github.com/MIABIS/miabis/wiki

4.14. Structural Data extension: Structural data

Module	Data Type
Structural Data extension	**Structural data**

Specification of content

Structural data comprises aggregated information on the individual, legally independent organizational units in a data integration center. This includes not only the medical services available but also staff numbers, technical equipment, bed count, case figures and much more, for all of the hospital's clinics and departments. This information can be found in the structured quality reports published pursuant to §137 of Volume V of the Social Security Statute Book[39]. These reports provide case figures of the diagnoses made (per ICD-1-GM) and the procedures carried out (per OPS) in each department where the number exceeds a predetermined threshold (data privacy).

Reason for assignment to the core data set

Structural data are of some interest in connection with simple queries such as feasibility checks, since they allow for an estimate of the upper limit to the number of existing patients with particular indications in the hospitals. For example, if a research hypothesis dealing with acute myeloblastic leukemia (C92.0) requires data on 1,000 patients, data integration centers that report having the corresponding population sizes or other appropriate conditions for collaboration can be selected.

Recommendations for structuring and coding

Since 2003, all hospitals have been required to prepare quality reports with the aim of providing patients with an overview of the type and scope of various medical services provided by the hospitals as well as several quality indicators (e.g., side-effects, mortality rates). Since 2008, quality reports have been prepared in a standardized (i.e., according to a pattern) XML format[40] pursuant to the guidelines set by Germany's Federal Joint Committee (GB-A). They consequently satisfy essential technical prerequisites for inclusion in the core data set, since they (1) are available for all university hospitals, (2) are in a common format with common semantics, and (3) have been in existence for about 10 years, which will reduce the work required for loading into research databases.

Recommendation for how to proceed

Having those in positions of authority at the DICs request the quarterly report data from all of the hospitals associated with a data integration center would be a practical step, as would transformation into the local DIC information model. Due to the manageable degree of complexity, this can occur as early as 2018. If coordinated, DIC-spanning information models were to become established, centrally developed and

[39] https://www.gesetze-im-internet.de/sgb_5/__137.html
[40] https://www.g-ba.de/institution/themenschwerpunkte/qualitaetssicherung/qualitaetsdaten/qualitaetsbericht/xml-daten/

maintained ETL components would also be conceivable. Stage: 1

4.15. Extension: Fees

Module	Data Type
Extension	**Fees**

Specification of content

The “Fees” table in the Sec21-data set contains information on the remuneration for cases in the hospital.

Reason for assignment to the core data set

The information on fees can contain valuable clues for health services research. Because of the legitimate business interests of the commercial entity that is a hospital, particular restrictions on availability are advised.

Recommendations for structuring and coding

The structuring and coding of a “Fees” table in the MII core data set can be based on the “Fees” table of the Sec21-data set.

Recommendation for how to proceed

For the MII core data set, a “Fees” table analogous to the corresponding table of the Sec21-data set should be planned in the DIC starting phase. Coordinated adjustments can be made as needed. Stage: 1

4.16. Extension: Cost Data

Module	Data Type
Cost Data extension	**Cost data (calculation for the InEK)**

Specification of content

The "Costs" table in the expanded Sec21-data set contains vectors with cost information for the cases in the hospital.

Reason for assignment to the core data set

The information on costs can contain valuable clues for health services research. A number of university hospitals participate in the annual calculation of treatment costs run by the InEK (Institution for the Hospital Remuneration System), which means that corresponding data are available.
Because of the legitimate business interests of the commercial entity that is a hospital, particular restrictions on availability are advised.

Recommendations for structuring and coding

The structuring and coding of a "Costs" table in the MII core data set can be based on the "Costs" table of the Sec21-data set. Formats and calculation standards are described in detail in the InEK's Calculation Handbook[41].

Recommendation for how to proceed

For the MII core data set, a "Costs" table analogous to the corresponding table of the Sec21-data set should be planned in the DIC starting phase.

Coordinated adjustments can be made as needed.

Stage: 1

[41] http://www.g-drg.de/Kalkulation2/DRG-Fallpauschalen_17b_KHG/Kalkulationshandbuch

5. Persons in Charge of Updating the Core Data Set

Modules	Data Types	In Charge
Basic module	Person	J. Schepers. F. Prasser
	Demographics	J. Schepers
	Case data	J. Schepers
	Diagnoses	J. Schepers
	Procedures	J. Schepers
	Lab results	M. Löbe, S.C. Semler
	Medication	T. Ganslandt, M. Löbe
Oncology extension	Tumor data as per the "ADT-GEKID"	T. Ganslandt
Extension, diagnostics	Pathology results	B. Böckmann
	Imaging findings	B. Böckmann, M. Boeker
Critical Care Medicine extension	PDMS data and high-resolution biosignal data	J. Ingenerf
OMICS extension	Genetic tests and sequencing	M. Boeker
Biobank extension	Biomaterial data	T. Ganslandt, S.C. Semler
Structural Data extension	Structural data	M. Löbe
Fees and Cost Data extension	Fees	J. Schepers
	Cost data (calculation for the InEK)	J. Schepers

Drafting Group Members and Contributors

Britta Böckmann
Martin Boeker
Karoline Buckow
Thomas Ganslandt (Leader, Core Data Set Drafting Group)
Josef Ingenerf
Matthias Löbe
Fabian Prasser
Josef Schepers
Sebastian C. Semler