PERSIMUNE Datawarehouse overview and catalogue

Version 1.0

Author: PERSIMUNE Datawarehouse group

This work is the culmination of many people working over many years. For a full list of those who have contributed, see appendix 1.

Updates since previous version.

First version. So all text is new.

Contents

1 Brief introduction

PERSIMUNE (www.persimune.dk) is a collaborative centre of excellence embedded within CHIP (www.chip.dk) and Rigshospitalet. The centre aims to identify novel risk factors for infectious outcomes in persons with immune dysfunction and develop and implement precision medicine strategies to improve these outcomes. To achieve this, PERSIMUNE invested in a data infrastructure, termed the PERSIMUNE Datawarehouse (PM DWH) that could connect the different hospital and registry databases relevant to studying the patient populations of interest. With the correct regulatory and/or ethics approval, data within the PM DWH can be extracted for research purposes. To aid in this process of providing data for research, this document provides an overview of the PM DWH; the structure, data sources, tables contained within and where to find further information. The information in this documentation is not intended to replace a proper project feasibility assessment, which can be conducted in collaboration with PERSIMUNE IT, but rather inform potential researchers on the data available within the PM DWH and to act as a compendium of knowledge related to these data.

Each section in this document is completed by those with active knowledge of a particular area of the PERSIMUNE Datawarehouse. As some areas (e.g. Microbiology) are more commonly used, these areas have more in depth descriptions and more documentation available than those areas that have not been so extensively used. The document is also constantly being updated and peer reviewed by the PERSIMUNE data warehouse group in order to ensure accuracy of existing data and update areas with less information. If you notice an error or would like to contribute based on your own knoweldge of a particular element outlined below, please contact RH-FP-PERSIMUNE <persimune.rigshospitalet@regionh.dk>.

2 PM DWH structure

Data within the PM DWH is organised into 'Domains'. These domains represent biologically or medically similar data. A single domain can contain data from one or multiple sources. The Ad-Hoc domain contains data that is either project specific or otherwise does not fit into another domain. Some data sources within the different domains are from national registries whereas others contain geographically limited data (e.g. either from Region Hovedstaden or Rigshospitalet only). Further, as some data sources have either ceased to exist or access to these data sources has stopped, data within each domain may also only cover a particular time period. These details, stratified by domain, are outlined in Table 1. For more detailed technical specifications of the PM DWH, please contact PERSIMUNE.

Table 1 Datawarehouse structure and time/geographic breakdown

Domain	Geography	IT-System/Data Source	Data Type (Time Frame)
Biobank	Rigshospitalet	PERSIMUNE Biobank	Biobank **(2016-2021)**
Bloodbank		LabWare / Blodbank	Blodtype **(2005-)**
			Transfusion **(2005-)**
Biochemistry	National	Labatorie Database(LDB)Laba (If new data, pushed via this system)	All kinds of Biochemistry Data **(1980)**
		LSP / MEDCOM	Biochemistry data **(2004-)**
		LABKA II	Biochemistry results directly from the lab production system **(2009-)**
	Rigshospitalet	KK Labka	A list of historical biochemistry analysis result. Contains a limited set of analysis. **(2005-2009)**
		VAEVSTYPELAB	TBNK **(2005-)**
			HLA **(2005-)**
Covid	Rigshospitalet	CATCH 1 (Testing) CATCH 2 (Hospitalized COVID patients at RH and RH Glostrup)	Covid-19 Monitoring & Testing **(2020-2021)** X-rays scores by MBrixia Vitals
Demography & Diagnoses	RegionH	CØK	Diagnoses, Procedures, Referral **(2007-2020)**
		Sundhedsdatabanken	
	National	FS	LPR3 **(1900 - 2022)**
		CØK	LPR 2 **(2005-2020)**

Imaging/diagnostics	RH (Rigshospitalet)	RIS/PACS	Imaging/diagnostics codes **(2005-)**
	HVH (Hvidovre)		
	BB (Bispebjerg)		
	FRB (Frederiksberg)		
Medication/ Ordination/ Admin	RegionH	EPM1	Administrative Data **(2005-2011)**
		EPM3	Administrative Data **(2012-2016)**
Microbiology	National	LSP / MEDCOM (including MiBa)	XRPT05 **(2004-)**
	Rigshospitalet	MADS	Microbiology results directly from the lab production system. **(2005-)**
Oncology	Rigshospitalet	PPAS	Information on oncology patients medication **(2005-)**
	Herlev	HOBS	Information on oncology patients medication **(2006-)**
Pathology	National	LSP (MEDCOM), Patobank	Pathology Data **(2005-)**
Radiation	Rigshospitalet	ARIA09 ARIA11 ARIADB001	Radiation Data **(2009-)**
		ECLIPSE	
	Herlev	ARIA	Radiation Data **(2007-)**
		ECLIPSE	Eclipse Data () To be received
Transplantation	Rigshospitalet	MATCH	Transplantation data **(2008-)**
	Herlev		
Vital Signs	Rigshospitalet	CØK / KISO	Information on vital signs as blood pressure, pulse, temperature, height, weight etc. **(2013-2016)**
		ICCA / ICCA_SIRS	Information on vital signs as blood pressure, pulse, temperature, height, weight etc. **(2007-2017)**
		ISCV	Information on vital signs as blood pressure, pulse, temperature, height, weight etc. **(2005-2016)**
Ad-Hoc	RegionH	PACS	Thorax Rtg (CATCH indlagte) **(2020-2022)**

		LYFO	Project Specific Data **(2005-2015)**
		Biobank	Leftovers, reference to data in Biobank A list of material stored in the KMA (clinical microbiology department at Rigshospitalet) "Biobank til fremtidig forskning i udvidet diagnostik og værtsrespons ved kompliceret sygdom" Leftover under biobank **(2010-2015)**
	National	Dansk Receptdatabase	Prescription Data **(2005-2016)**
		FS (FSEID3121)	LPR (FSEID3121) **(2005-)**
			CAR (FSEID3121) **(2005-)**
			CPR (FSEID3121) **(2005-)**
			DAR (FSEID3121) **(2005-)**
		Bio- og Genombank	Metadata **(2010-2015)**

3 Data import, QA and XDEP tables

Data exists, and can therefore be requested for research purposes, in three formats:

- Import tables – contains tables directly from the data source. For more details on these, please see documentation provided by each data source outlined in Table 2 or contact PERSIMUNE IT to conduct a feasibility assessment.
- QA tables – tables that have had automatic PERSIMUNE cleaning and QA rules applied. Not all variables from import tables are included in QA tables and multiple import tables may be combined into a single QA table.
- XDEP tables – tables that are structured for research or clinical application purposes.

Each domain may contain multiple import, QA and XDEP tables. For an overview of all import, QA and XDEP tables for each of the PM DWH domains, please see Table 2 and the text below.

For more details on the individual domains and the tables within, see the Domain specific subsections of this document.

Table 2 Overview of data structure within the PERSIMUNE data warehouse

Domain	IT-System/ Data Source	Import Table	QA_Table (Persimune_Data)

Biobank	PERSIMUNE Biobank	Data imported in tbl_ADM_IMPORT_PM_BB this table is truncated and the data moved to tbl_ADM_PM_BB	X
Biobank	Rigshospitalet CancerBiobank center	DCB_EDTAfuldblod DCB_EDTAplasma DCB_serum	X
Bloodbank	LabWare / Blodbank	tbl_TEMP_IMPORT_Blodbank tbl_TEMP_IMPORT_Blodbank_Transfusions tbl_TEMP_IMPORT_Blodbank_Bloodtype tbl_TEMP_IMPORT_Blodbank_TEG tbl_TEMP_IMPORT_Blodbank_Multiplate tbl_TEMP_IMPORT_Blodbank_Noak_Pradaxa_TEG tbl_TEMP_IMPORT_Blodbank_Recipient_komplikationer tbl_TEMP_IMPORT_Blodbank_BIVIRKNINGER_PATIENT_DONOR tbl_TEMP_IMPORT_Blodbank_Mater_CPR_RH_Profylakse_og_Antenatal	X
Biochemistry	Labatorie Database(LDB) Laba (If new data, pushed via this system)	tbl_IMPORT_LABA_Lab_Resultat_XRPT01??? NULL in the tblAPP_Settings_SQL_ImportData	X
Biochemistry	LSP / MEDCOM (To request CPR's)	tbl_IMPORT_Medcom_Lab_Resultat_XRPT01 tbl_IMPORT_Medcom_Lab_Logistic	tbl_QA_Biochemistry
Biochemistry	LABKA II	tbl_IMPORT_LABKA_REGIONH_SHADOW	tbl_QA_Biochemistry
Biochemistry	KK Labka	tbl_labka	tbl_QA_Biochemistry
Biochemistry	VAEVSTYPELAB	temp_TBNK	tbl_QA_Biochemistry

Biochemistry	VAEVSTYPELAB	temp_HLA temp_HLA_schema temp_HLA_NGS temp_DuraClone temp_DuraClone_mapping	tbl_QA_Biochemistry
Covid	CATCH	tbl_TEMP_IMPORT_CORONA_REGIONH tbl_TEMP_IMPORT_CORONA_REGIONH_LOG_FUND NULL tbl_TEMP_IMPORT_LABKA_REGIONH	
Demography & Diagnoses	CØK / Sundhedsdatabanken	SDB_V_HENVIS SDB_V_KAT SDB_V_KON SDB_V_OPHO SDB_V_PAS SDB_V_Patient_master SDB_V_REKV SDB_V_SKSDIA (Only [PERSIMUNE_DATA].[dbo].[tbl_ADM_Import_SourceTables]) SDB_V_PROC	tbl_QA_Diagnoses (SDB_V_SKSDIA) tbl_QA_Procedures (SDB_V_PROC) tbl_QA_Referral (SDB_V_HENVIS)
Demography & Diagnoses	LPR	Import_ALL	X
Demography & Diagnoses	CØK / LPR	LPR_All_XXXX (Diag, Bes,….)	X
Imaging/ diagnostics	RIS/PACS	tbl_ADM_IMPORT_RIS_REGIONH	tbl_QA_Radiology

Medication/ Ordination/ Admin	EPM1	epm1_ordination	tbl_QA_AdministrationEPM1 tbl_QA_Ordination_EPM1
Medication/ Ordination/ Admin	EPM3	EPM_V_MedicinOrdinationer EPM_V_MedicinDoseringer EPM_V_MedicinAdministrationer EPM_V_MedicinOrdination_ENSU	tbl_QA_Ordination(EPM_V_MedicinOrdinationer) tbl_QA_Dosing(EPM_V_MedicinDoseringer) tbl_QA_MedicineAdministration(EPM_V_MedicinAdministrationer) tbl_QA_OrdinationENSU (EPM_V_MedicinOrdination_ENSU)
Microbiology	MADS	tbl_TEMP_IMPORT_MADS_REGIONH tbl_TEMP_IMPORT_MADS_REGIONH_LOG_FUND	tbl_QA_MicrobiologyAnalysis
Microbiology	LSP (MEDCOM)	All tables starting with tbl_IMPORT_Medcom_Lab ? 1.tbl_IMPORT_Medcom_Lab_Resultat_XRPT05 2 tbl_IMPORT_Medcom_Lab_Logistic 3. tbl_IMPORT_Medcom_Lab_Resultat_XRPT05_Investigation	**tbl_QA_MicrobiologyCultures** **tbl_QA_MicrobiologyCulturesResistance** **tbl_QA_MicrobiologyMicroscopy** tbl_QA_MicrobiologyAnalysis
Oncology	PPAS	tblOncology	tbl_QA_Oncology_v2
Oncology	HOBS	tblOncology	tbl_QA_Oncology_v2
Pathology	Patobank / LSP (Medcom)	tbl_IMPORT_Medcom_Lab_Resultat_XRPT04 tbl_IMPORT_Medcom_Lab_Logistic tbl_IMPORT_Medcom_Lab_Resultat_XRPT04_CodedFormat	
Radiation	ARIADB001 ARIA09 ARIA11	tbl_RADIATION_TREATMENT tbl_RADIATION_TREATMENT_DIAGNOSIS	tbl_QA_RadiationTreatment tbl_QA_RadiationTreatmentDiagnosis

Transplantation	MATCH	All tables starting with tbl_IMPORT_Medcom_Lab_Logistic	X
Vital Signs	ICCA / ICCA_SIRS	temp_import_icca_data (IMPORT_ICCA_DATA)	tbl_QA_ICCA tbl_QA_ICCA_SIRS
Vital Signs	ISCV	import_iscv_data	X
Vital Signs	KISO / CØK	KISO_V_Klinisk_data KISO_V_Overordnet	tbl_QA_Clinical
Ad-Hoc	PACS	tbl_IMPORT_CORONA_REGIONH	X
Ad-Hoc	???	tblEKG	
Ad-Hoc	Dansk Receptdatabase	dansk_receptdatabase_2017AUG25	tbl_QA_DanishPrescriptionDatabase
Ad-Hoc	FS / LPR	_FSID3121_XXXX	
Ad-Hoc	Bio- og Genombank	[dbo].[tblRegionernesBioGenonBank_20160530]	
Ad-Hoc	Biobank	tbl_BULK_KMA_LeftOvers	tbl_QA_TYPE_MIBA_Categorical
Ad-Hoc	LYFO	tbl_ADM_IMPORT_LYFO_2016-11-14_JDL_Nov_2016_ORIGINAL	

4 XDEP tables

XDEP tables are tables that have been structured for research or clinical application purposes. Data within an individual XDEP table may come from multiple QA tables and from multiple domains. Their naming nomenclature and structure is based on the HIV Cohorts Data Exchange Protocol (HICDEP) https://hicdep.org/. For the LAB and LAB_VIRO XDEP tables, PERSIMUNE researchers have grouped together variables that have different names and/or analysis codes in the source data, but mean the same thing from a research perspective. These mostly relate to Biochemistry and virological analyses. and are a result of data being reported from multiple timeperiods and from multiple laboratories. The grouped variables and the analysis names, codes and source included in the groupings are outlined further in the XDEP folder in the corresponding zip file to this catalogue. If you wish to utilise the PM DWH to

access biochemistry or study viral phenotypes, we highly recommend reviewing these prior to starting. It is possible that variables have been changed or added since the cleaning last took place. As such, it is always best practice to review the last time each variable of interest was reviewed and cleaned by PERSIMUNE researchers or staff (link to this documentation) and, if in doubt, to update the data using the relevant QA table.

For more details on the individual XDEP tables or to learn more, please contact PERSIMUNE.

Table index	Description	Coverage
BAS	Basis information such as demographics	
CEP	Type and date of clinical events	
CEP_DIAG	Type and date of clinical events for diagnosis	Capital Region
LAB	Biochemistry laboratory tests	National
LAB_CLINICAL	Clinical information (KISO). Height, weight, blood pressure etc.	Capital Region
LAB_CULT	Microbiology - Cultures	National
LAB_MICRO	Microbiology - Microscopy	National
LAB_VIRO	Microbiology - Virus analyses	National
LTFU	Lost to follow-up - death & drop out	
MED	Medicin from EPM1 and EPM3	Capital Region
REFER		Capital Region
RAD	Radiology (Rigshospitalet)	Rigshospitalet
PAT	Pathology (national)	National
CEP_PROC	Type and date of clinical events for procedures	Capital Region

4.1 CEP tables – Data enrichment using clinical classification algorithms

The CEP table contains clinical classification algorithms. These are variables which combine different data variables to indicate a clinical diagnosis or classification which is otherwise not available or not reliable in the data received form the various sources. These algorithms have been generated as part of PERSIMUNE research projects and have been validated by clinicians working within those projects. As the

data sources or data variables . If one would like to use these variables outside the validated period or for clinical purposes, then additional validation may be required.

Febrile Neutropaenia

Developed by Theis Aagaard

The published description of the FN outcome is as follows:

"FN was defined as a blood culture or death within three days of a neutrophil count $<0.5x10^9$/L or a leucocyte count $\leq 2.0x10^9$/L if neutrophils were not measured[1]. Data on temperature measurements were not routinely available before 2014 and were only available for the Capital Region of Denmark and hence a blood culture was used as a measure of clinical suspicion of infection. This definition has been shown to be in good concordance with a narrower guidelines-based definition of FN: neutropenia $<0.5x10^9$/L and fever ≥38 degrees Celsius[1]."*(ref: CSR-FENCE study,).*

Data dictionary for table CEP for the FN outcome

CEP_ID = FN.

CEP_SPEC = FN_BC_NEU if blood cultures and neutrophils are used, FN_BC_LEUK if blood cultures and leukocytes are used, FN_DEATH_NEU if death and neutrophils are used, and FN_DEATH_LEUK if death and leukocytes are used.

CEP_D = Date of blood culture sample or death

Datasource = supplier of data for blood culture or death (MEDCOM supplies data from MIBA on blood cultures, CPR supplies data from the Civil Registration Service on death dates).

Datasource2 = supplier of data for biochemistry values (LABKA supplies data from Rigshospitalet and the Capital Region of Denmark, MEDCOM supplies nationwide data from multiple data systems)

Please note that death data is no longer recieved from CPR. Although death dates are available within the PERSIMUNE DWH, to utilise this phenotype correctly, a project would need to clean the algorithm to map death data to the new source.

ICU stay

More details can be found in the XDEP folder

Cancer diagnosis

Stata and SQL code to generate a cancer diagnosis at a particular time point – e.g at time of radiation therapy or chemotherapy initiation. Calculated using data form the PM DWH. Developed by Cynthia Terrones Campos. Contact PERSIMUNE for more details regarding this phenotype.

Infections

Documentation and description pending

Failure to cancer treatment (cancer relapse)

Documentation and description pending

Post-transplant Diabetes

Documentation and description pending

Charlson CoMorbidity Index

The classification of relevant ICD-10 diagnoses for the individual comorbidities is performed according to the algorithm developed by Quan et al[1]. The original index was presented by Charlson[2] in 1987 with 19 variables, which were collapsed to 17 variables by Deyo et al.[3] in a study where administrative data where used to calculate the Charlson Comorbidity Index score. The Charlson Comorbidity Index was updated by Quan et al.[4] In a large international study of 7 million patients published in 2011, reflecting the weights of individual comorbidities on risk of mortality in a modern health-care setting. The Charlson Comorbidity Index has been validated in data from the Danish National Patient Registry (NPR) where the positive predictive values of the diagnoses found in the NPR was found to be 98% with a range of 88-100%[5].

In PERSIMUNE the score is calculated from all A (primary)- and B(secondary)-diagnoses from both inpatient and outpatient visits on or before indexdate. In PERSIMUNE we have assessed the score when we only use inpatient diagnoses and found this to be erroneous in a Danish setting where many patients are seen in an outpatient setting. Presently, we deliver two versions of the Charlson Comorbidity Index score, one where all diagnoses from inpatient and outpatient diagnoses are used (CCI_2) and one where we only use the diagnoses given one year before the indexdate (CCI_4).

The indexdate has to be defined by each project.

More details in the PERSIMUNE QA Rapport for Charlson Comorbidity Index Score - English version

Positive bloodstream infections

Documentation and description pending

Additional documentation for CEP and XDEP related data

See the XDEP folder in the corresponding zip file.

5 Biobank Domain

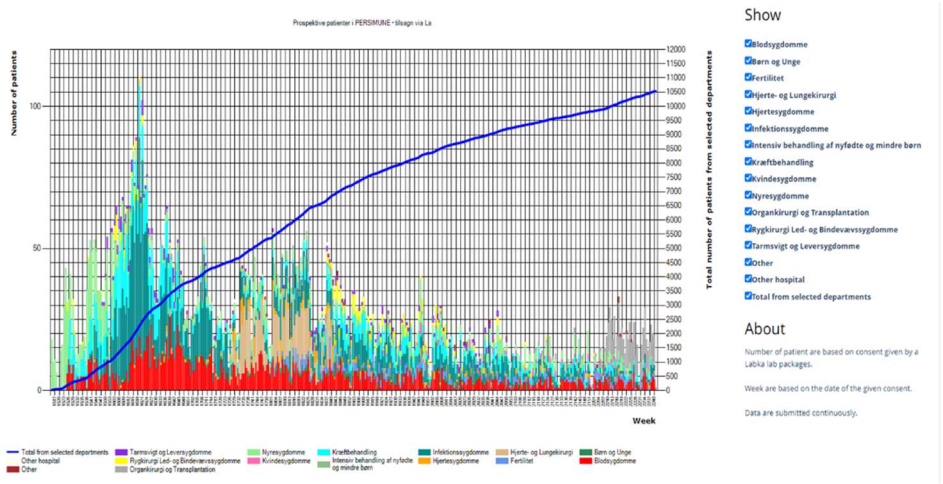
Brief description of domain:

The Biobank Domain contains information on biological samples collected as part of the PERSIMUNE biobank. The PERSIMUNE biobank is approved by the Danish Data Protection agency (RH-2015-04, I-suite 03605) and all patients with samples collected as part of the PERSIMUNE biobank have consented to their samples being used for future research. Although access to these samples requires additional ethical and PERSIMUNE approval.

There are currently over 10000 patients who have consented to the PERSIMUNE biobank from various departments collaborting with PERSIMUNE (see figure below). Plasma and whole blood samples are the most commonly collected samples, while there is also a large collection of feacal samples . For more details regarding the population level details on samples collected or to better understand the approvals required to access these samples, please contact PERSIMUNE.

Samples contained within the biobank include

- Plasma
- Fuldblod (Whole Blood)
- Feaces
- Ekspektorat (Expectorate / Sputum)
- Bronchoalveolar lavage
- Blood spot
- DNA extracted from whole blood
- Saliva



WHen utilising samples from the biobank, it is important to note that in some cases the collection date for a sample is not available. In these cases, one can use Freeze Date (the time and date the sample was placed in the freezer) to get an idea of when the sample was taken. As samples are processed the same day as they are collected, the freeze date should reflect the collection date.

Link to further documentation:

See biobanks.dk to search diagnoses and samples from within the PERSIMUNE and other Danish biobanks

6 Bloodbank Domain

Brief description of domain:

To be updated in next version

Link to further documentation:

7 Biochemistry Domain

Brief description of domain: The biochemistry domain contains information on biochemical analyses performed at clinical biochemical departments at Danish hospitals as well as other laboratories within Denmark. The information available in the domain includes, but is not limited to, analysis codes, analysis names, analysis responsible and performing laboratory, analysis times and dates, analysis material, test results, reference intervals, and their respective units of measurement. Data in the biochemistry domain is sourced from LABKA, Medcom, and Vævstypelab and is requested by researchers from one singular QA table. Medcom is a national source of biochemistry data, LABKA is from the Capital Region and Vævstypelab is the Rigshospitalet tissue typing lab. Continuous data cleaning efforts were previously put in place to group appropriate analyses despite inter- and intra-laboratory discrepancies in analysis codes and names; however, documentation on standardized data cleaning in this domain may need to be updated depending on the time period a researcher wishes to study. There are also some biochemistry and viral analyses in the Blodbanken, but these are outlined in the documentation for that domain.

Link to further documentation:

See an overview of the Biochemistry grouping and cleaning efforts in the XDEP folder of the zipped files.

https://www.medcom.dk/

8 COVID

Brief description of domain:

Several data sources are available for COVID-19, but they are governed by different regulatory approvals. In-house treatment databases CATCH1 (SARS-CoV-2 PCR testing of out- and in-patients) and CATCH2 (In-patients from Rigshospitalet and Rigshospitalet Glostrup) follow the PERSIMUNE regulatory approvals on treatment databases. Data on all COVID-19 patients at hospitals in Region Hovedstaden and Region Sjælland is available through collaboration with COVIMUN, but cannot be accessed without prior approval from the PIs of COVIMUNE (Sisse Ostrowski and Carsten Utoft Niemann).

CATCH 1:

Data from treatment database used for booking of SARS-CoV-2 PCR testing samples., that was used for out-patient testing in the early COVID-19 pandemic (Mar-Apr 2020) and later for in-patient testing, ambulatory visit booking from a range of departments at Rigshospitalet, as well as testing efforts for travelers to Greenland until early 2022.

*CATCH 2:*Data collected from sundhedsplatformen on patients admitted to either COVID-19 specific departments, intensive care departments, and infectious disease department at Rigshospitalet and Rigshospitalet Glostrup.

Contains 1-2 snapshots a day of data on admitted patients including vitals, prescribed medications, microbiological samples, biochemistry, admission status and death dates. Also a large number of x-rays from this patient population was scored with the MBrixia score by radiologists at Rigshospitalet.

COVIMUN:

Very comprehensive dataset extracted from backend of Sundhedsplatformen (SP) containing a wide range of variables on COVID-19 patients treated in The Capital Region of Denmark and Region Zealand.

DATA includes the following variables and more:
Patient identifier: Pseudonymized identifiers.
Hospital visits: in-patient and out-patient stays
Biochemistry: Missing values.
Medication: All medicine prescribed and administered during the hospitalization..
Microbiology: SARS-CoV-2 PCR,
Vitals: Blood pressure, respiratory support, respiratory frequency, oxygen saturation, pulse etc.
Death status and date: No cause of death available.

Link to further documentation:

9 Demography and Diagnoses

Brief description of domain: This domain contains information relating to Demography and clinical diagnoses, as well as other key information such as department and hospital contact/admissions. Data within this domain comes from Sunhedsdatabanken and LPR. More recent data from LPR3 is sourced from Forskerservice and can only be accessed through the PERSIMUNE data protection approval. As these data come in from multiple sources and cover different time periods, some cleaning is required to fully utilise data within this domain.

Link to further documentation:

10 Imaging/Diagnostics

Brief description of domain:

To be updated

Link to further documentation:

11 Medication/Ordination/Admin

Data sources: EPM1 and EPM3

Brief description of domain: The electronic clinical medication system, EPM, was used to prescribe medication in the hospitals in the capital region. This system consists of two main versions, EPM1 which was the first system (used in 2005-2011) and was later replaced by EPM3 (2011-2016). During the year of 2016 the EPM3 system has been replaced by Sundhedsplatformen (SP), with different hospitals changing to this system at different time-points. As our patient cohort is based on patients treated at Rigshospitalet, this means that data from EPM3 stops from November 2016 onwards (unless the ordination was made prior to this date and had a stop date after).

Data from EPM1 and EPM3 are extracted to PERSIMUNE via the central data distribution platform located in Hillerød. An

Overall, the data include patient identifier, prescribed drug names and ATC codes, start and end dates of the prescription, prescribed drug form and way, and the prescribed drug doses.

Of note, the PERSIMUNE Medication Domain does not contain data on in-hospital medicines (asides from some information on Oncology medication; see Oncology domain) past 2016.

Further data documentation:

See corresponding folder in zip file

12 Microbiology Domain

Brief description of domain:

Nationwide microbiological tests and results are available through MiBa from 2010 and onwards. MiBa is a national clinical system launched 1/1/2010, initiated by the Statens Serums Institute (SSI). All Danish clinical microbiological departments started to upload the local laboratory data into MiBa automatically from 1/1/2010. The upload of SSI data to MiBa on test performed in the central SSI laboratory was initiated 11/3/2013.

Microbiology data is organised in different tables depending on type of investigation/test performed; 1) Cultures, 2) Resistance, 3) Microscopy, 4) Molecular biological analyses (e.g. PCR, ELISA). Due to the heterogeneity of sources providing data to MiBa (several clinical microbiological departments from all over Denmark) the microbiological data including naming of sample material, investigations and results are also somewhat heterogeneous and data cleaning is required. Some infectious diseases have been investigated in previous PERSIMUNE research projects and the data regarding certain infections have therefor been cleaned and validated to some extent. These cleaned mostly relate to viral infections and are further outlined in the LAB_VIRO Grouped and Coding tables that can be found in the XDEP folder of the corresponding ZIP file. Information on previous data cleaning performed and guidance in further data cleaning advised can be provided on request by researchers with experience in the field. However, as these data come from many labs and involve many different measurements (e.g. cultures, PCRs, resistance assays, serology etc), there is a potential for variable names to change over time and prior to starting analysis these should be checked.

Link to further documentation:

Link to MiBa docs
https://miba.ssi.dk/mikrobiologidatabase
https://miba.ssi.dk/service/english

Also see the corresponding Zip file of the

13 Oncology

Brief description of domain: The data in the QA_oncology table comes from the PPAS chemotherapy ordering system used at the Departments of Oncology at Rigshospitalet, Herlev, and Hillerød. Generally, data on the chemo drugs are data on drugs delivered from the pharmacy meant for each individual patient. However, if the treatment is postponed after chemo has been delivered to the department, the

drugs will be thrown out but data still be in the QA_oncology table. Most often a new requisition for the chemo will appear a few days later which is actually a double. Only intravenous chemotherapy is recorded in the database. Trastuzumab (Herceptin) is not always recorded in the db as the pharmacy does not prepare this specific drug. Accordingly, all oral drugs are also not in the db.

Link to further documentation:

14 Pathology

Brief description of domain: Data from the national pathological database "Patobank" containing patient identifier, date of sampling, who requested the investigation and who performed it. Data on the material investigated, the type of investigations performed, and the results are available via SNOMED coding provided by the pathologist in a uniform manner used nationwide allowing easy identification of relevant sample types, investigations performed and results of interest.

Link to further documentation:

SNOMED code book avaialbe: https://www.patobank.dk/

15 Radiation Oncology

Brief description of domain: We have extracted relevant information from Rigshospitalet's and Herlev´s radiotherapy record and verify database, ARIA®, regarding radiotherapy characteristics provided to patients (i.e. radiation plan, number of fractions, dose per fraction, etc). Quality assurance processes have been identified and removed. Dosimetry data were only available from 2009 due to a change in systems. We currently only store dosimetry data for the body structure; however, data from other delineated structures could be obtained.

Link to further documentation: Varian® documentation at www.myvarian.com, upon registration.

16 Transplantation

Brief description of domain: The Transplantation domain contains data from the MATCH clinical application (https://www.chip.dk/Clinical-programs/MATCH). Tables within this domain are presented in the XDEP format and only for patients transplanted (either solid organ or haematopoetic) at Rigshospitalet. Data within this domain includes basic and demographic information, diagnosis, clinical procedures (including transplant type), grouped laboratory values, grouped viral measurements, serology, pathology and lost to follow-up/ death data.

In addition to these data, the CLASS project aims to accurately classify the cause of death for patients within the MATCH program (https://www.chip.dk/Clinical-programs/MATCH/CLASS). As such, clinically validated causes of death are also available for patients within MATCH.

Link to further documentation: For an overview of the different tables and an SOP for use of MATCH data, please see MATCH_STC_SOP_v1_5_2018SEP27.

17 Vital Signs

Brief description: Contains data on vital signs from the Rigshospitalet KISO database. Only available until 2016, after which vital sign information is contained within SP. These data are not available within the PM DWH

18 Ad Hoc

Brief description: The Ad Hoc domain contains data that has either not yet been added to another domain or does not fit neatly into another domain. It also contains project specific data that can be uploaded to the PM DWH, combined with PM data and delivered in a secure pseudoanymised format for research. The project specific data will not be shared within this data catalogue.

The Ad Hoc domain currently also contains data from Forskerservice. These data are, however, only available when using the PERSIMUNE research approval. (=data from CPR-registry; Cancer-registry; Cause of Death-registry (DAR); LPR & LPR-3).

Link to further documentation: For more information on data contained within the Ad Hoc domain, please contact PERSIMUNE IT.

For information on Forskerservice data, see the Ad Hoc folder in the corresponding zipped folder.

19 Appendix 1

Michael Asger Andersen