DALYCARE Database Views

Technical Documentation and Dependency Analysis

---

1. Overview

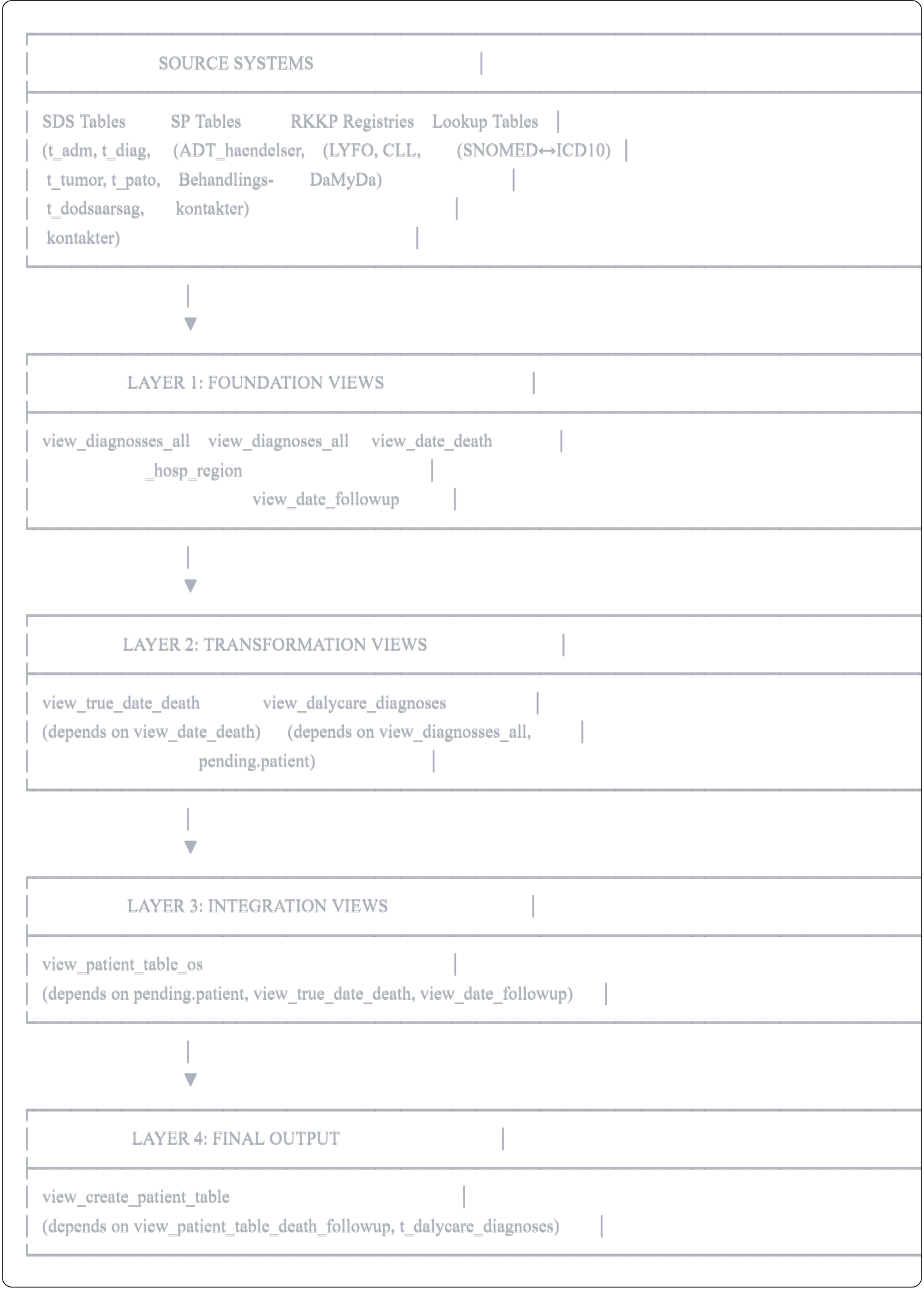
This document provides comprehensive technical documentation for the DALYCARE database views. These views are part of a clinical data system designed to manage hematological malignancy patient data, integrating information from multiple Danish health registries including SDS (Sundhedsdatastyrelsen), RKKP (Regionernes Kliniske Kvalitetsudviklingsprogram), and hospital administrative systems.

The views form a hierarchical structure where foundation-level views aggregate raw data from source tables, intermediate views transform and filter this data, and final views produce patient-level analytical datasets suitable for research and clinical quality assessment.

---

2. View Dependency Graph

The following diagram illustrates the dependency relationships between views. Arrows indicate data flow from source to dependent view. Views are organized in layers based on their position in the dependency hierarchy.



Dependency Summary Table

Layer	View Name	Depends On
Layer 1 (Foundation)	view_diagnosses_all	Base tables: SDS_, *RKKP_*, SP_*, lookup tables
Layer 1	view_diagnoses_all_hosp_region	Base tables: SDS_, *RKKP_*, lookup tables
Layer 1	view_date_death	patient, SP_ADT_haendelser, SDS_t_dodsaarsag, RKKP_*
Layer 1	view_date_followup	patient, SP_ADT_haendelser, RKKP_*, SDS_t_dodsaarsag
Layer 2 (Transform)	view_true_date_death	view_date_death
Layer 2	view_dalycare_diagnoses	view_diagnosses_all, patient
Layer 3 (Aggregate)	view_patient_table_os	patient, view_true_date_death, view_date_followup
Layer 4 (Final)	view_create_patient_table	view_patient_table_death_followup, t_dalycare_diagnoses

3. Detailed View Documentation

3.1 Diagnosis Data Pipeline

The diagnosis views form the core data pipeline for capturing and filtering hematological malignancy diagnoses from multiple source systems. These views are tightly coupled and should be considered as a unified pipeline.

3.1.1 view_diagnosses_all

Schema: pending

Purpose: Consolidates diagnosis data from all available Danish health registries into a unified format. This is the foundational view for all diagnosis-related queries.

Data Sources:

The view aggregates data from multiple source systems using UNION operations to create a comprehensive diagnosis dataset. SDS (Sundhedsdatastyrelsen) provides data from several tables including t_adm for admission diagnoses, t_diag for detailed diagnosis records, t_tumor for cancer registry entries, t_dodsaarsag_2 for cause of death diagnoses, kontakter for contact-based diagnoses, and pato for pathology findings. SP (Sundhedsplatformen) contributes through ADT_haendelser for ADT event diagnoses and BehandlingskontakterOgDiagnoser for treatment contact diagnoses. RKKP registries provide disease-specific data from LYFO for lymphoma patients, CLL for chronic lymphocytic leukemia, and DaMyDa for multiple myeloma. Additionally, lookup tables provide SNOMED to ICD-10 code mappings that enable integration of pathology data.

Output Columns:

Column	Type	Description
patientid	integer	Unique patient identifier
date_diagnosis	date	Date when diagnosis was recorded
diagnosis	text	ICD-10 diagnosis code (Danish variant with D prefix)
tablename	text	Source table name for data lineage
datasource	text	Source system: SDS, SP, RKKP, or PATO
priority	integer	Data quality priority (1=PATO highest, 2=RKKP, 3=SDS, 4=SP)

Key Filtering Logic:

The view applies several important filters to ensure data quality. It excludes CLL diagnosis code DC910, which is handled by a separate dedicated registry. Only diagnoses from 2002-01-01 onwards are included, as this represents the beginning of reliable electronic data capture. Future dates are filtered out by requiring that the diagnosis date must be less than or equal to the current date, which prevents data entry errors from propagating to analytical datasets.

Priority System:

The priority column reflects the relative reliability of each data source for hematological malignancy diagnoses. Pathology data (PATO) receives priority 1 as it represents confirmed histological diagnoses. RKKP clinical quality registries receive priority 2 due to their specialized curation for hematological conditions. SDS national registry data receives priority 3 as general administrative data. SP hospital system data receives priority 4 as it may contain preliminary or rule-out diagnoses.

3.1.2 view_diagnoses_all_hosp_region

Schema: pending

Purpose: Extended version of view_diagnosses_all that includes hospital and region information for geographic analysis of diagnoses.

Additional Columns:

Column	Type	Description
hospital	integer	Hospital code (c_sgh) where diagnosis was recorded
region	integer	Danish health region identifier (1-5)

This view follows identical logic to view_diagnosses_all but preserves geographic identifiers that would otherwise be lost in the union operations. The hospital code uses the Danish SGH (Sygehus) coding system, while the region field maps to the five Danish health regions: Region Hovedstaden (Capital Region), Region Sjælland (Zealand), Region Syddanmark (Southern Denmark), Region Midtjylland (Central Denmark), and Region Nordjylland (North Denmark).

Geographic analysis is essential for understanding regional variations in diagnosis patterns, access to specialized care, and outcomes across different parts of Denmark's healthcare system.

3.1.3 view_dalycare_diagnoses

Schema: public

Purpose: Filters diagnoses to include only adult patients (age ≥ 18 at diagnosis) with specific hematological malignancy codes. This is the primary cohort definition view for the DALYCARE study.

Dependencies:

This view depends on pending.view_diagnosses_all for the consolidated diagnosis data and pending.patient for date of birth information needed to calculate age at diagnosis.

Included Diagnosis Codes (ICD-10):

The view includes patients with diagnoses in the following categories. The DC81-DC91 range covers Hodgkin lymphoma through myeloid leukemia, encompassing the major hematological malignancies. Specific codes DC951, DC957, and DC959 capture acute leukemia variants not covered in the main range. Codes DD472, DD472A, DD472B, and DD479B identify MGUS (Monoclonal Gammopathy of Undetermined Significance) and related plasma cell conditions. Code DE858A captures amyloidosis related to plasma cell disorders, which often co-occurs with myeloma.

Exclusion Logic:

The view implements sophisticated exclusion logic using an EXCEPT clause. This removes patients who had a pediatric diagnosis (age < 18) from a non-PATO source and later received an adult diagnosis. This approach ensures that pediatric-onset cases are not incorrectly classified as adult-onset based on follow-up diagnoses recorded after the patient turned 18.

The rationale for excluding PATO from this rule is that pathology diagnoses are definitive histological confirmations. If a patient has a pathology-confirmed diagnosis as an adult, this represents a true adult diagnosis regardless of any earlier administrative entries that might have been recorded when the patient was a minor.

3.2 Survival and Follow-up Pipeline

These views form a tightly integrated pipeline for determining patient vital status, death dates, and follow-up endpoints. They are essential for survival analysis and should be modified together when changes are required.

3.2.1 view_date_death

Schema: pending

Purpose: Aggregates death date information from all available sources into a single view with source-specific columns. Provides the raw data layer for death date determination.

Death Date Sources:

Column	Source Description
sp_date_death	Sundhedsplatformen ADT events (most current hospital data)
sds_date_death	SDS death certificate registry (official civil registration)
rkkp_lyfo_date_dead	RKKP Lymphoma registry (DD/MM/YYYY format)
rkkp_cll_date_death	RKKP CLL registry (DD/MM/YYYY format)
rkkp_damyda_date_death	RKKP Myeloma registry (DD/MM/YYYY format)

Data Validation:

RKKP dates are stored as text strings in DD/MM/YYYY format and require validation before conversion. The view checks that the string length is exactly 10 characters (representing a valid DD/MM/YYYY date). Invalid dates, such as empty strings or malformed entries, are set to NULL to prevent conversion errors in downstream processing.

Preliminary true_date_death Calculation:

This view includes a preliminary true_date_death column using a CASE statement that prioritizes sources in the following order: SP, SDS, DaMyDa, CLL. However, note that the downstream view_true_date_death uses a different priority order that better reflects the reliability of sources for hematological patients.

3.2.2 view_true_date_death

Schema: pending

Purpose: Determines the authoritative death date by applying a priority cascade across all death date sources. Handles date format conversions and provides a single true_date_death column.

Priority Cascade (highest to lowest):

The view applies the following priority order when multiple death dates are available for a patient. First priority goes to RKKP DaMyDa (Myeloma registry), followed by RKKP CLL (CLL registry), then RKKP LYFO (Lymphoma registry), then SP (Sundhedsplatformen), and finally SDS (Death certificate registry) as the lowest priority.

Rationale for Priority Order:

The RKKP registries take precedence because they are specialized clinical quality registries with dedicated data managers who actively curate patient records. For hematological malignancy patients, these registries typically have the most current and accurate death information because they maintain ongoing follow-up of their patient populations.

SP data comes next because it reflects real-time hospital information, though it may have delays in recording deaths that occur outside the hospital system. SDS death certificate data, while official, may lag due to administrative processing times and is therefore given lowest priority for timeliness, though it remains the authoritative source for legal purposes.

Important Note:

The priority order differs between view_date_death (which prioritizes SP first) and view_true_date_death (which prioritizes RKKP first). The view_true_date_death priority order is considered more appropriate for the DALYCARE hematological malignancy cohort.

3.2.3 view_date_followup

Schema: pending

Purpose: Collects the most recent follow-up date from each data source for living patients (those without a death date in the respective source system).

Follow-up Date Sources:

Column	Description
date_sp_followup	Maximum death date from SP (used as administrative follow-up cutoff)
date_lyfo_followup	CPR update date from RKKP LYFO for patients without death date
date_damyda_followup	CPR update date from RKKP DaMyDa for patients without death date
date_cll_followup	CPR update date from RKKP CLL for patients without death date
date_sds_followup	Maximum status date from SDS death registry

Follow-up Date Interpretation:

For living patients, the follow-up date represents the date through which we have confirmed the patient was alive. The CPR_Opdat_dt (CPR Update Date) fields from RKKP registries indicate when the patient's vital status was last verified against the Danish Civil Registration System (CPR). The SP follow-up uses the maximum death date in the system as an administrative cutoff, indicating that patients not recorded as dead by this date were presumed alive.

3.3 Patient Analytical Tables

These views combine patient demographics with survival endpoints to create analysis-ready datasets. They are the primary entry points for research queries.

3.3.1 view_patient_table_os

Schema: pending

Purpose: Creates a patient-level table with overall survival (OS) endpoints suitable for survival analysis. Combines demographics, death dates, and follow-up information.

Output Columns:

Column	Type	Description
patientid	integer	Unique patient identifier
sex	text	Patient sex
date_birth	date	Date of birth
date_death	date	Death date (NULL if patient is alive)
status	integer	Vital status indicator: 0 = alive/censored, 1 = deceased
date_followup	date	Last known alive date (NULL if patient is deceased)

Survival Analysis Usage:

This view is designed specifically for survival analysis workflows. For time-to-event analysis, researchers should use date_death as the event date and date_followup as the censoring date. The status column provides the event indicator required by survival analysis methods, where 1 indicates the patient experienced the event (death) and 0 indicates the patient was censored (still alive at last follow-up).

Time-to-event can be calculated from a reference date (typically the diagnosis date from view_dalycare_diagnoses) to either date_death (if status=1) or date_followup (if status=0). This structure is compatible with standard survival analysis packages in R (survival package), Python (lifelines), and SAS (PROC LIFETEST, PROC PHREG).

Status Calculation Logic:

The status column is derived using a simple CASE statement: if true_date_death is not NULL, status equals 1; otherwise, status equals 0. The date_max_followup is calculated as the GREATEST of all available follow-up dates, but only for patients where true_date_death is NULL.

3.3.2 view_create_patient_table

Schema: pending

Purpose: Final patient table that filters to only include patients present in the DALYCARE cohort (those with diagnoses in t_dalycare_diagnoses). Provides a unified date_death_fu column for simplified survival analysis.

Key Features:

This view restricts the patient population to DALYCARE cohort members only, ensuring that all patients in the output have at least one qualifying hematological malignancy diagnosis. It combines the death date and follow-up date into a single date_death_fu column, which simplifies survival analysis by providing a single endpoint date regardless of vital status. The status indicator (0/1) is preserved to distinguish events from censored observations.

Output Columns:

Column	Type	Description
patientid	integer	Unique patient identifier
sex	text	Patient sex
date_birth	date	Date of birth
status	integer	Vital status: 0 = alive, 1 = deceased
date_death_fu	date	Combined endpoint: death date if deceased, follow-up date if alive

Dependency Note:

This view depends on view_patient_table_death_followup, which was not included in the provided SQL files. Before using this view, verify that this dependency exists and is properly maintained. The t_dalycare_diagnoses table (not a view) should be populated from view_dalycare_diagnoses.

---

4. Complete Data Flow

Source Systems

SDS (Sundhedsdatastyrelsen): The Danish Health Data Authority provides national health data including hospital admissions (t_adm), detailed diagnoses (t_diag), cancer registry entries (t_tumor), pathology reports (pato), death certificates (t_dodsaarsag_2), and outpatient contacts (kontakter). This represents the comprehensive national administrative health data infrastructure.

SP (Sundhedsplatformen): The shared electronic health record system used in the Capital Region and Region Zealand provides real-time hospital data through ADT (Admission, Discharge, Transfer) events and treatment contact records. SP data is typically more current than SDS but covers only patients treated in the two regions using this system.

RKKP (Regionernes Kliniske Kvalitetsudviklingsprogram): The Danish Clinical Registries provide specialized quality databases for specific conditions. LYFO covers lymphoma patients, CLL covers chronic lymphocytic leukemia, and DaMyDa covers multiple myeloma (Danish Myeloma Database). These registries include detailed clinical information curated by specialists.

Lookup Tables: Code mapping tables, particularly SNOMED to ICD-10 mappings, enable integration of pathology data (which uses SNOMED codes) with the ICD-10-based diagnosis infrastructure.

Processing Pipeline

Stage 1 - Aggregation Layer: The foundation views (view_diagnosses_all, view_diagnoses_all_hosp_region, view_date_death, view_date_followup) consolidate data from multiple sources into standardized formats. These views handle the complexity of different data structures, date formats, and coding systems across source systems.

Stage 2 - Transformation Layer: The transformation views apply business rules to select authoritative values and define cohorts. view_true_date_death applies the priority cascade to determine the single best death date for each patient. view_dalycare_diagnoses applies the age and diagnosis code criteria to define the adult hematological malignancy cohort.

Stage 3 - Integration Layer: view_patient_table_os joins patient demographics with survival endpoints, creating a unified patient-level analytical dataset with all variables needed for survival analysis.

Stage 4 - Presentation Layer: view_create_patient_table provides the final analysis-ready dataset filtered to the DALYCARE cohort, with simplified column structure optimized for common analytical workflows.

---

5. Technical Notes

5.1 Date Format Handling

RKKP registries store dates in DD/MM/YYYY format as text strings, reflecting the Danish date convention. The views use PostgreSQL's to_date() function with the 'DD/MM/YYYY' format string for conversion. Invalid dates (strings that are not exactly 10 characters, indicating incomplete or malformed entries) are filtered out by setting them to NULL before conversion. This prevents conversion errors that would otherwise cause query failures.

Example conversion pattern:

```sql
to_date(rkkp_cll_date_death, 'DD/MM/YYYY')
```

5.2 Type Casting Considerations

Several joins in these views cast patientid between integer and double precision types. This pattern appears in constructs such as:

```sql
ON a.patientid = b.patientid::double precision
```

This suggests potential data type inconsistencies in source tables, where some tables store patientid as integer and others as numeric/double precision. While functional, this implicit type conversion has performance implications and should ideally be addressed at the ETL level by standardizing the patientid data type across all tables.

5.3 Known Spelling Variant

The view `view_diagnosses_all` contains a spelling variant (double 's' in 'diagnosses'). This should be maintained for backward compatibility with existing queries, reports, and application code that reference this view name. Any future refactoring should include appropriate aliases or synonyms to prevent breaking changes.

5.4 Age Calculation Method

Age at diagnosis is calculated using the formula:

```sql
((date_diagnosis::date - date_birth::date)::numeric / 365.25)::real
```

This approach divides the number of days between dates by 365.25 (accounting for leap years) to obtain age in years. The result is cast to real (single-precision floating point). For the adult filter (age ≥ 18), this method is sufficiently accurate, though it may differ by a few days from exact calendar age calculations.

5.5 EXCEPT Clause Logic

The view_dalycare_diagnoses uses an EXCEPT clause to implement the pediatric exclusion logic:

```sql
SELECT ... FROM adult_diagnoses
EXCEPT
SELECT ... FROM (pediatric_diagnoses JOIN adult_diagnoses ON patient AND diagnosis)
```

This removes any adult diagnosis record that has a matching pediatric diagnosis for the same patient and same diagnosis code from a non-PATO source. The EXCEPT operation returns rows from the first query that do not appear in the second query, effectively filtering out adult records for patients whose condition was first diagnosed in childhood.

5.6 Permissions Model

Views are owned by the `casfre` user with SELECT permissions granted to various roles:

Role	Access Level	Description
casfre	ALL	View owner with full control
cll_lab	SELECT	Read access for CLL research team
cube_user	SELECT	Read access for reporting tools
importread	SELECT	Read access for ETL monitoring
importuser	ALL	Full access for ETL processes
mikbrs_r	ALL	Full access for specific user
maintenance	ALL	Full access for system maintenance

When modifying views, ensure appropriate permissions are maintained using GRANT statements after the view definition.

---

6. Appendix: View Creation Order

When recreating the database views, they must be created in dependency order. The following sequence ensures that all dependencies are satisfied:

1. **Base tables must exist:** pending.patient, all SDS_* tables, all RKKP_* tables, SP_* tables, lookup tables
2. **Layer 1 views (can be created in parallel):**
    - pending.view_diagnosses_all
    - pending.view_diagnoses_all_hosp_region
    - pending.view_date_death
    - pending.view_date_followup
3. **Layer 2 views (require Layer 1):**
    - pending.view_true_date_death (requires view_date_death)
    - public.view_dalycare_diagnoses (requires view_diagnosses_all)
4. **Layer 3 views (require Layer 2):**
    - pending.view_patient_table_os (requires view_true_date_death, view_date_followup)
5. **Layer 4 views (require Layer 3):**
    - pending.view_create_patient_table (requires view_patient_table_death_followup)

Document generated from SQL view analysis. Last updated: January 2026.