Data management at SIMSAM MEB node

Åsa Eck

Databas

- 12 register
- All data ligger i tabeller i en Oracle databas, version 11.2.
- Data har laddats in till Oracle via SAS. Loggarna är sparade som kommentarer i SAS programmet.

Databas

- Schema MGR_SOCMOB_RAW innehåller all rå-data.
- Schema MGR_SOCMOB innehåller alla tabeller som används av forskare.
- Schema MGR_SOCMOB_PREPARE används för att rena tabellerna och inte för att lagra data.
- För att göra sökningar snabbare och effektivare är index skapade i Oracle på de kolumner som efterfrågas mest.

Rättigheter

- DM-ansvarig för projektet (av MEB kallad DBA), PI (Principal Investigator) och forskare har alla konto i Oracle
- Ett antal roller är upplagda i Oracle för att hantera rättigheterna i SIMSAM-databasen (rollerna har skapats av en Oracle DBA)
- PI bestämmer vem som ska ha tillgång till vad i databasen och när dessa rättigheter ska tas bort
- DBA är ansvarig att ge och ta bort medlemskap i dessa roller
- DM och PI har tillgång till alla register

Rättigheter (forts)

- Forskare har endast tillgång till de register som de behöver för sin forskning.
- DM har läs och skriv-rättighet till alla register
- PI och forskare har endast läs-rättighet
- Forskare programmerar i SAS och loggar därifrån in i Oracle för att få tillgång till data.
- Cirka 35 forskare har jobbat med registren

Alla register – Rening av data

- Reningen av data har skett genom att köra SQL-kommandon och SAS-program.
- Alla register består av två tabeller. Dessa tabeller heter RegisternamnÅr_CLEAN och RegisternamnÅr_INVALID, t.ex.
  OUTPATIENT0107_CLEAN, OUTPATIENT0107_INVALID, OUTPATIENT0610_CLEAN, OUTPATIENT0610_INVALID
- CLEAN-tabellerna innehåller de rader som har renats och har giltig information.
- INVALID-tabellerna innehåller de rader som har tagits bort vid rening av data och alltså innehåller någon slags ogiltig eller otydbar information.

Alla register – Rening av data (forts)

- Följande rening har gjorts:

- Alla text-kolumner med LOPNR omvandlas till integer-format

- En tabell som heter INVALID_LOPNR är skapad som innehåller alla ogiltiga LOPNR, dvs. alla personer med återanvända personnummer, ogiltiga personnummer och dubbla personnummer. Även personer som vi inte vet när de invandrade/utvandrade finns med i denna tabell.
  Denna tabell används endast vid reningen av data och endast DBA har tillgång till den.

Alla register – Rening av data (forts)

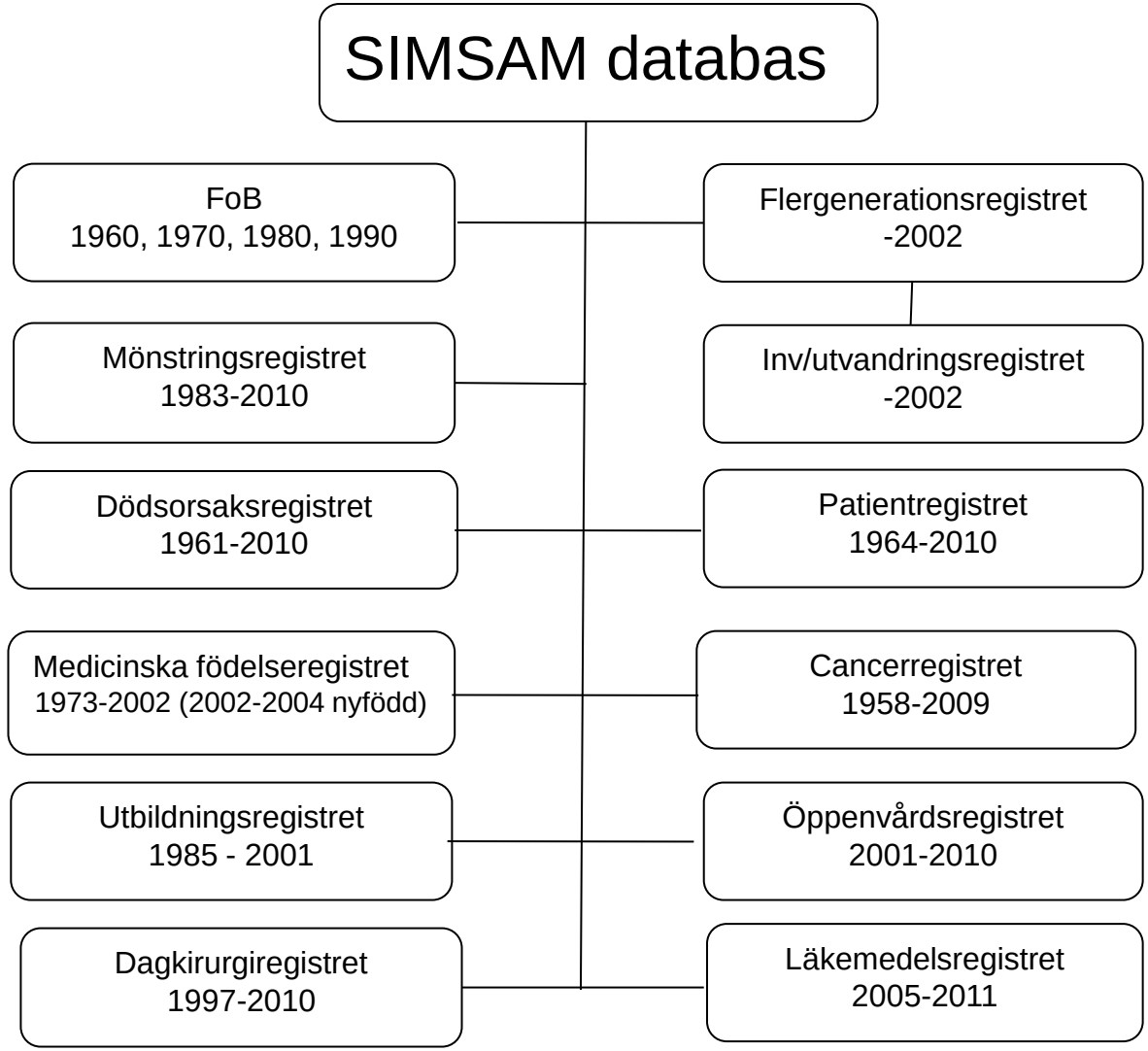
- Alla personer i tabell INVALID_LOPNR är flyttade från _CLEAN till _INVALID-tabellerna.
- En extra kolumn är skapad i stort sett alla _INVALID-tabellerna med en flagga som indikerar varför raden är ogiltig.
- För flergenerationstabellerna sätts ogiltiga LOPNR till NULL i _CLEAN-tabellen.

Alla register – Rening av data (forts)

- Felaktiga datum är härledda till en ny kolumn vars namn börjar på X_ (t.ex. X_DIADAT om original heter DIADAT) enligt följande:
  - ogiltig månad härleds till 0701
  - om dag inte är mellan 01-31 blir den härledd till 15
  - om dag=31 på en 30-dagars månad blir den härledd till 30
  - om felaktig dag för sista februari blir den härledd till korrekt dag

- Alla text-kolumner med datum härleds till datum-format. Namnet på kolumnen avslutas med _DATE, (t.ex. DIADAT_DATE)

- En cleaning-rapport dokumenterar all rening som har gjorts.


SIMSAM databas
FoB
1960, 1970, 1980, 1990
Mönstringsregistret
1983-2010
Dödsorsaksregistret
1961-2010
Medicinska födelseregistret
1973-2002 (2002-2004 nyfödd)
Utbildningsregistret
1985 - 2001
Dagkirurgiregistret
1997-2010
Flergenerationsregistret
-2002
Inv/utvandringsregistret
-2002
Patientregistret
1964-2010
Cancerregistret
1958-2009
Öppenvårdsregistret
2001-2010
Läkemedelsregistret
2005-2011

INDIVID

- Populationen i tabell Individ utgörs av alla svensk-födda individer i flergenerationsregistret mellan år 1932-2002 och deras föräldrar. I populationen ingår även alla som varit med i någon av folkbokföringsräkningarna som genomfördes 1960-1990.

- Individ_Clean
  13.598.327 rader

- Individ_Invalid
  6.797 rader

Karolinska Institutet

MGR (flergenerationsregistret)

- Flergenerationsregistret till och med år 2002
- BioAdForaldrar_Clean (Barn-Föräldrar relation)
  7.739.033 rader
- Barn_Clean (Individ-Barn relation - Biologiska)
  14.998.364 rader
- AdBarn_Clean (Individ-Adoptivbarn relation)
  219.841 rader
- Helsyskon_Clean (Helsyskon relation)
  12.528.608 rader

MGR (forts)

- Halvsyskon_Clean (Halvsyskon relation)
  3.618.058 rader
- AdSyskon_Clean (Adoptivsyskon relation)
  53.942 rader
- BioAdSyskon_Clean (Biologisk-Adoptivsyskon relation)
  26.577 rader

Invandring/Utvandringsregistret

- Invandring/Utvandringsregistret till och med år 2002
- InvUtv_Clean
  3.235.401 rader

MFR (medicinska födelseregistret)

- Medicinska födelseregistret mellan år 1973–2002
- MFR_CLEAN där tabellen blivit normaliserad och diagnoserna utflyttade till nya tabeller.
  2.990.067 rader
- För att spara utrymme har diagnoserna flyttats ut till egna tabeller då de senare diagnoserna ofta är tomma.

MFR (forts)

- MFR_BDIAG_ORDER_CLEAN (en rad per BDIAG)
  3.483.449 rader
- MFR_MDIAG_ORDER_CLEAN (en rad per MDIAG)
  4.909.146 rader
- MFR_FLOP_ORDER_CLEAN (en rad per FLOP)
  1.568.116 rader
- MFR_GDIAG_ORDER_CLEAN (en rad per GDIAG)
  451.326 rader

MFR (forts)

- MFR_DF (dödfödda barn) där tabellen är normaliserad på samma sätt som MFR_CLEAN. Eftersom dödfödda barn inte får något LOPNR härleds LOPNR_DF genom att ta moderns MOR_LOPNR_x (där x står för löpnummer i dödfödda barn)
  13.030 rader

- MFR_NY (barn födda mellan 2002-2004) där tabellen är normaliserad på samma sätt som MFR_CLEAN. Eftersom dessa barn inte ingår i populationen har de inte något LOPNR. Därför härleds LOPNR_NY genom att ta moderns MOR_LOPNR_x (där x står för löpnummer i barn födda mellan 2002-2004)
  190.982 rader

FOB

- Folk och Bostadsräkningarna mellan år 1960-1990
- FOB60_Clean – år 1960
  7.467.069 rader
- FOB70_Clean – år 1970
  8.073.434 rader
- FOB80_Clean – år 1980
  8.318.187 rader
- FOB90_Clean – år 1990
  8.584.951 rader

Utbildningsregistret

- Utbildningsregistret mellan år 1985-2001
- EducationReg_Clean
  147.621.269 rader

Cancerregistret

- Cancerregistret mellan år 1958-2009
- Cancer5809_Clean
  2.278.007 rader

Dödsorsaksregistret

- Dödsorsaksregisret mellan år 1952-2010
- DeathReg6110_Clean – mellan år 1961-2010
  4.440.775 rader
- DeathReg5260_Clean – mellan år 1952–1960
  78.176 rader

Patientregistret

- Patientregistret mellan år 1964-2010
- InPatient6405
  49.270.542 rader
- InPatient6407
  49.880.769 rader
- InPatient0610
  7.085.340 rader
- Inpatient6410_Clean (Inpatient_6405 + Inpatient_0610)
  54.131.742 rader

Öppenvårdsregistret
Dagkirurgiregistret

- Öppenvårdsregistret mellan 2001-2010
- Dagkirurgiregistret mellan 1997-2010
- Outpatient0107_Clean
  47.201.551 rader
- Outpatient0610_Clean
  39.972.510 rader
- Daysurgery9707_Clean
  8.439.020 rader

Öppenvårdsregistret
Dagkirurgiregistret (forts)

- Eftersom ovanstående tre register överlappar varandra ska en flagga skapas i varje tabell som indikerar om vårdtillfället förekommer i mer än ett av registren. Dock skall inga rader rensas bort eftersom det är svårt att bedöma om det är dubbletter eller inte.

Läkemedelsregistret

- Läkemedelsregistret mellan juni 2005–juni 2011
- DrugPrescription05_Clean
  34.498.024 rader
- DrugPrescription06_Clean
  71.603.233 rader
- DrugPrescription07_Clean
  73.772.066 rader
- DrugPrescription08_Clean
  76.048.983 rader

Läkemedelsregistret (forts)

- DrugPrescription09_Clean
  80.094.610 rader
- DrugPrescription10_Clean
  80.891.639 rader
- DrugPrescription11_Clean
  52.770.863 rader

Mönstringsregistret

- Mönstringsregistret mellan år 1983–2010

- Conscription8397_Clean
  753.644 rader
- Conscription9701_Clean
  213.700 rader
- Conscription0208_Clean
  120.147 rader
- Conscription0910_Clean
  16.874 rader

Standarder/guidelines på MEB

- MEB_guideline_Documentation_and_Archiving

  Hur arbetet ska struktureras under ett pågående projekt (t.ex. en bra folder-struktur, en readme-fil under varje folder som beskriver vad foldern innehåller).
  Dokumentation av sitt arbete
  Hur man arkiverar avslutat projekt.

- MEB_policy_DataStandards

  Standard på namn
  Förkortningar i variabelnamn
  Koder

Att göra allmänt

- Skapa ett dokument med översättning av alla tabell och kolumnnamn till engelska. Även förklaring vad dom innehåller.

  Några översättningar finns redan i SIMSAM, inom andra projekt på MEB (Institutionen för Medicinsk Epidemiologi och Biostatistik) och från myndigheter som levererar data.

  Något som kanske kan göras gemensamt över noderna?

Frågor?