folkehelseinstituttet

Health Registries for Research

Norway

Rapport 2/2017

Delprosjekt

«KPR-analyser: Etablering av forskningsmiljø for KUHR- og NPR-data i TSD»

Dokumentplassering:	F:\Forskningsprosjekter\PDB 2010 - KPR analyseprosjekte_\Prosjektdokumentasjon\Rapporter
Forfattere:	Øystein Jonasson, Marta Ebbing, Mette C. Tollånes
Godkjent av/dato:	Marta Ebbing, 20.2.2017

Endringsoversikt

Versjon	Dato	Hvem/status	Beskrivelse/endringer
0.1	14.7.2016	Marta Ebbing	Mal til utfylling.
0.2	31.2.2017	Marta Ebbing	Videre bearbeidet mal til utfylling.
0.3	3.2.2017	Øystein Jonasson	Utkast til gjennomlesning ved Marta Ebbing.
0.4	6.2.2017	Marta Ebbing	Tilbakemeldinger til Øystein Jonasson.
0.5	8.2.2017	Øystein Jonasson	Flettet inn avsnitt fra tidligere versjoner.
0.9	13.2.2017	Marta Ebbing	Ferdigstilling, klart for gjennomlesning ved forskere.
0.98	19.2.2017	Marta Ebbing	Tilbakemeldinger fra Mette Tollånes, Anja Ariansen og Stein Emil Vollset tatt inn i dokumentet. Mangler bare noen få fakta, som Øystein Jonasson sjekker opp.
1.0	20.2.2017	Øystein Jonasson	Endelig versjon.

Innholdsfortegnelse

1 Innledning

Dette delprosjektet er knyttet til infrastrukturprosjektet «Health Registries for Research», finansiert av Norges forskningsråd, og Folkehelseinstituttets bidrag til dette arbeidet.

Hensikten med prosjektet var å evaluere bruken av Tjenester for Sensitive Data (TSD) ved Universitet i Oslo (UiO) med hensyn på analyse av større mengder helsedata i ulike forsknings- og helseanalyseprosjekter. I den anledning var det naturlig å benytte KPR-analyseprosjektet for å støtte opp om etableringen av Kommunalt pasient- og brukerregister (KPR) som pilot for bruken av TSD ettersom det var store datamengder og en prosjektgruppe som var geografisk adspredt. En dataforvalter fra Avdeling for helsedataforvaltning og –analyse (HDFA) var allerede allokert til prosjektet og kunne fungere som teknisk prosjekt koordinator mot TSD.

Denne rapporten beskriver erfaringene fra KPR-analyseprosjektet med hensyn til tilrettelegging av store datamengder for forskning og analyse, samt etablering og bruken av forsknings- og analysemiljø i TSD.

2 Nærmere beskrivelse av oppgaver, funn og problemer

Dette kapittelet tar for seg bruk av tid og ressurser for de forskjellige stadiene i analysen.

2.1 Mottak, kobling og tilrettelegging av data for analyser

For KPR-analyseprosjektet ble det benyttet uttrekk av data fra Norsk pasientregister (NPR) og system for Kontroll og utbetaling av helserefusjon (KUHR), som beskrevet i tabellen under. Tilgang til data var bestilt av Helsedirektoratet som prosjekteier og databehandlingsansvarlig. Direkte personidentifiserbare data ble levert fra hver av kildene med hver sin koblingsnøkkel, som så skulle kobles mot en populasjon bestående av alle personer som var registrert som «bosatte» i Det sentrale folkeregisteret (DSF) i løpet av perioden 2006-2014. Detaljert innhold i datakildene er ikke relevant for denne rapporten og vil ikke bli diskutert nærmere.

Kilde	Årganger	Antall hendelser (mill.)	Filstørrelse (GB)
NPR (Område 4 og 5)	2008-2014	55,6	4,8
KUHR	2006-2014	292,0	23,9

I forbindelse av mottak av data ble det tilrettelagt på «sikker sone» ved Folkehelseinstituttet i Bergen for behandling av direkte personidentifiserbare data, både fil-område for lagring av mottatte filer samt database-skjema for behandling av data. Tilgang var gitt kun til dataforvaltere ved HDFA og IT-drift personell. Innlasting av data til databasen ble utført i henhold til standard rutiner ved HDFA, men grunnet store datamengder var dette mer tidkrevende enn ved vanlige innlastingsjobber.

Første steg i tilretteleggingen var å definere populasjonsgrunnlaget for analysen. Dette ble definert som alle personer som hadde potensiale for å være mottakere av helsetjenester i løpet av analyseperioden. Dette ble konkretisert som alle personer født senest i 2014, som var enten registrert som bosatt i DSF per 31.12.2014 eller død/utvandret i løpet av 2006 eller senere. Hvert individ i denne populasjonen ble gitt et unikt prosjekt-spesifikt løpenummer, som ble koblet til relevante persondata fra DSF, KUHR og NPR for individet.

Aller helst skulle denne populasjonsdefinisjonen ha forekommet før uthenting av data fra andre registre, slik at koblingen mellom individ og løpenummer (kalt koblings-fil eller -nøkkel) kan gjenbrukes av alle registre som deltar i koblingen. Denne teknikken betegnes ofte «distribuert kobling» og gjør at mottagende forsker/analytiker enkelt kan sammenstille data fra flere kilder ved å kun se på løpenummeret, og således ikke trenger tilgang på direkte personidentifiserende informasjon, som fødselsnummer. Ettersom dette ikke var tilfelle i KPR-analyseprosjektet, var uttrekkene fra KUHR og NPR koblet til hver sin unike løpenummerserie, slik at hvert individ kunne ha opptil tre distinkte løpenummer. For å kunne koble de mottatte datafilene med prosjektpopulasjonen var det følgelig nødvendig å gjøre et dobbelt oppslag for hver av de over 300 millioner hendelsene (konsultasjon i KUHR eller episode i NPR):

I prosjektet var det fire ulike sykdomsgrupper som skulle analyseres, definert ved gitte diagnosekoder. For å sikre uniform utvelgelse ble det opprettet en kalkulert variabel for hver av de fire sykdomsgruppene. For hver hendelse ble disse så satt til 0 eller 1 avhengig om diagnosene assosiert med hendelsen tilfredsstilte de oppgitte kriteriene. En annen fordel med denne

fremgangsmåten var at de ble enklere for forskere å filtrere hvilke hendelser som var relevant for hver sykdomsgruppe. Tilsvarende ble det på individ-nivå summert opp hvor mange konsultasjoner/episoder som var registrert i hver av datakildene for hver av de fire sykdomsgruppene, noe som forenklet arbeidet med å finne hendelsene som var ikke direkte relatert til sykdommen. Eksempelvis, skal man se på helsetjenesteforbruket til personer med diabetes, så vil ikke alle hendelser være kodet med diabetes direkte.

Siste del av tilretteleggingen var å trekke ut, kryptere og overføre datafilene til forskningsområdene ved Folkehelseinstituttet og TSD. Totalt var det tre forskjellige datasett som ble eksportert: DSF-data med relaterte variabler samt koblede og tilrettelagte utgaver av helsedata fra KUHR og NPR både som årgangs-filer og total-fil. Underveis ble det diskutert om det skulle tilrettelegges egne datafiler for hver av de fire sykdomsgruppene, men dette ble ikke ansett som nødvendig. Grunnet de store datamengdene tok det over en dag å pakke datafilene i krypterte arkiv.

Ekstern ressurs	**FHI ressurs**	**Oppgave**	**Merknad**	**FHI tidsbruk (timer)**
Dataforvaltere Helfo/Hdir (Vegard Håvik)	Dataforvalter (Jon Gunnar Tufta)	Administrere mottak av data fra KUHR fra Helfo		4
Dataforvalter NPR (Lillian Leistad)	Dataforvalter (Jon Gunnar Tufta)	Administrere mottak av data fra NPR	Tekniske problemer med overføring av uttrekk	12
	Dataforvalter (Øystein Jonasson)	Innlasting av data, definere populasjon, utføre kobling av datasett, uttrekk og overføring.		140
	IT-ressurs (Yngve Klakegg)	Oppsett og administrasjon av database-skjema	Det ble opprettet nye databaseskjema for person og helsedata med egen krypteringsnøkkel	12
	IT-ressurs (Oleksandr Kholosha)	Tilrettelegging av DSF-kopi med ny krypteringsnøkkel	DSF kopien i MFR miljøet ble gjenbrukt med ny krypteringsnøkkel	8

2.2 Etablering av forskningsmiljø i Tjeneste for sensitive data (TSD) ved Universitetet i Oslo

Prosessen med å opprette forskningsmiljø (sikret virtuell maskin) i TSD ble påbegynt i starten av mai 2016, før alle dataleveransene var mottatt. Normal saksbehandlingstid for prosjektsøknader hos TSD er to uker, men grunnet helligdagene i mai måned tok det en ekstra uke før forskningsmiljøet var opprettet. Tilgang til forskningsmiljøet på TSD er kontrollert med to-faktors autentisering, hvor passord og QR-kode til token generator sendes i adskilt rekommandert post, som ble mottatt en uke etter oppretting av forskningsmiljøet.

Ettersom ikke alle prosjektdeltagere var oppgitt i søknaden måtte tilgang for disse bestilles etter at prosjektadministrators tilgang var i orden, noe som medførte en ny to ukers saksbehandling og en ukes postgang før forskere hadde tilgang til forskningsmiljøet. Ved fremtidige bestillinger kan denne forsinkelsen unngås ved at tilgang for alle deltagere bestilles sammen med forskningsmiljøet. Ettersom opprettingen av TSD gikk parallelt med tilrettelegging av forskningsfilene var det ikke en vesentlig forsinkelse, men begge prosesser ble noe forsinket av at det var samme fagressurs tilordnet begge oppgavene.

Det gikk også med noe tid på tilrettelegging av kommunikasjon mellom Folkehelseinstituttet og TSD (åpning av porter i brannmuren) og installasjon av programvare for tilkobling til TSD (VMWare Horizon). Dette er forsinkelser som vi antar ikke er relevant for fremtidige prosjekter (gitt at TSD ikke endrer sine påloggingsrutiner).

I første omgang benyttet KPR-analyseprosjektet hardware-ressursene som fulgte med et standard forskningsmiljø i TSD, men det ble tidlig klart at prosjektet ville kreve mer ressurser, både hardware og software. Vi kjøpte derfor både ekstra prosessor (6 CPU) og ekstra lagringsplass (61 GiB RAM) samt statistikk-programmene SPSS, SAS og STATA. På samme tidspunkt ble det klart at HRR arbeidspakke 9, ledet av Universitetet i Bergen, investerte i hardware-ressurser (8 CPU og 64 GiB RAM) hos TSD[1], og prosjektet kunne benytte noen av disse. Totalt tok det ca. to uker å få på plass utvidete ressurser, og i denne perioden var analysekapasiteten noe redusert.

Prosjektet opplevde kun ett større teknisk problem, som skyldtes at brannmur-reglene hos Folkehelseinstituttet ble tilbakestilt, slik at tilkoblingen mot TSD ikke var tilgjengelig i en periode. Ut over dette var tilbakemeldingen fra forskerne som jobbet på TSD at løsningen var lett tilgjengelig fra både kontor og hjem, og at ytelsen var upåklagelig.

[1] Prosjektnummer p170-win01

Nedenfor er en tabell med oversikt over ressurser, oppgaver og tidsbruk samt fakturerte beløp i forbindelse med etableringen av forskningsmiljøet i Folkehelseinstituttets forskningsserver og i TSD.

Ansvarlig ressurs	**Annen ressurs**	**Oppgave**	**Tidsbruk FHI (timer)**	**Beløp fakturert FHI (NOK)**
Prosjektleder Marta Ebbing	Forsker Siri Håberg, jurist Ragnhild Holst	Utforme analyseprotokoll, skaffe behandlingsgrunnlag for tilgang til data, søke om data fra ulike dataforvaltere, utforme og signere databehandleravtale med Helsedirektoratet.[2]	80	0
Prosjektleder Marta Ebbing	Forsker Siri Håberg, dataforvalter Yngve Pedersen	Opprette prosjekt i Prosjektdatabasen (PDB) og lagringsområde med tilgangsbegrensning til analysegruppen på F:\Forskningsprosjekter\PDB 2010 - KPR analyseprosjekte_\[3]	1	0
Dataforvalter Øystein Jonasson	Prosjektleder Marta Ebbing	Anskaffe forskningsmiljø i TSD i henhold til rammeavtale mellom Folkehelseinstituttet og UiO med noe utvidete ressurser	2	22 650,00
Dataforvalter Øystein Jonasson	IT-Service	Teknisk tilrettelegging for kommunikasjon mot TSD	8	0
Dataforvalter Øystein Jonasson		Bestille bruker- og programvaretilgang i TSD	2	0
Dataforvalter Øystein Jonasson	Forsker Stein Emil Vollset, IT-sjef Roger Schäffer	Bestille ytterligere utvidede ressurser (8 CPU, 64 GiB RAM) i TSD via HRR arbeidspakke 9 ved Universitetet i Bergen	6	0[4]
Dataforvalter Øystein Jonasson		Oppfølging og støtte til prosjektet	16	0
Dataforvalter Øystein Jonasson		Uttrekk og overføring av koblede data til TSD	6	0

[2] All tilgang til forskningsmiljø ved TSD forutsetter hjemmel for behandling av sensitive data, som REK-godkjenning, konsesjon fra Datatilsynet eller annet grunnlag. Protokollutforming er en forutsetning for alle slike godkjenninger.

[3] For løsning for behandling av sensitive data i Folkehelseinstituttets egne sikre løsning.

[4] HRR har investert i et cluster hos TSD til en samlet pris av NOK 400 000. Prosjektet fikk benytte deler av dette clusteret til en besparelse på 22 650.

3 Oppsummering og anbefalinger

3.1 Evaluering av TSD

Terskelen for å komme i gang med TSD opplevdes nokså høy for forskerne. Uten tidligere erfaring med konseptet opplevdes det som mange nye skritt å gå før man var inne og fikk jobbet som normalt. Dette førte til at to av fire forskere i prosjektet jobbet mot Folkehelseinstituttets interne løsning for sensitive data (F:\Forskningsprosjekter, PDB-nummererte mapper), heller enn TSD. Likevel må skriftlig veiledning sies å ha vært adekvat, og når man først var kommet i gang opplevdes løsningen som strømlinjeformet og funksjonell. Den var enkel å bruke, godt tilgjengelig/portabel og stabil. Det var også en åpenbar fordel at man ikke hadde behov for en kraftig lokal maskin for å gjøre tyngre beregninger.

Den største utfordringen kom da forskerne etter en oppdatering hos TSD opplevde at de ikke fikk koblet seg til tjenesten på nærmere tre uker. Etter en del undersøkelser ved IT-service hos FHI, viste seg at den tekniske feilen lå hos Folkehelseinstituttet og ikke hos TSD. Dette ville trolig blitt avdekket raskere dersom Folkehelseinstituttet hadde fått raskere og bedretilbakemeldinger fra TSD support underveis.

Et annet negativt aspekt med TSD er at lisenser for statistikk-programmer må betales for hver bruker per prosjekt, fordi forskermiljøene i TSD er organisert som separate installasjoner.

For et forskningsprosjekt som skal analysere allerede innsamlete data eller bruke nettbasert innsamling av egne data, kan TSD anbefales som forskningsmiljø. Av økonomiske hensyn er det en fordel om man kan benytte seg av programvare uten lisenskostnader, eksempelvis R for statistiske beregninger.

3.2 Planlegging og bestillerkompetanse

Flere av de store problemene til prosjektet kan direkte knyttes til mangelfull planlegging og bestillerkompetanse.

Tilgang til TSD for forskere ble forsinket med nesten en måned fordi dataforvalter hos Folkehelseinstituttet i første omgang kun bestilte tilgang til seg selv, og ikke til alle prosjektdeltagere da bestillingen om forskningsmiljø ble sendt. Etterbestillinger ble behandlet som en ny bestilling, med tilsvarende saksbehandlingstid. Dette illustrerer hvor viktig det er å sende inn en fullstendig bestilling i første omgang.

Bruk av separate koblingsnøkler for KUHR og NPR (i stedet for distribuert kobling ved bruk av DSF som populasjonsgrunnlag) medførte betraktelig merarbeid for å koble data. Prosjektgruppen anbefaler at forskere med erfaring fra ønsket datatype involveres på et tidlig stadium, blant annet for å forsikre at en entydig prosjektpopulasjon kan defineres. For prosjekter som benytter store datamengder er det også fordelaktig å involvere en dedikert dataforvalter på et tidlig stadium for å vurdere databehandlingsstrategi.

3.3 Formidling av erfaringer med dataforvaltning i KPR-analyseprosjektet

For å bidra til at våre erfaringer med planlegging og gjennomføring av KPR-analyseprosjektet, ble prosjektet presentert for øvrige medarbeidere ved Folkehelseinstituttet ved flere anledninger;

- Under «Tall som teller» ved Folkehelseinstituttet i Bergen 29. april og 11. november 2016
- Under forskermøte ved Folkehelseinstituttet i Bergen 9. mai 2016
- Under møte med Kunnskapssenteret i Folkehelseinstituttet 4. oktober 2016
- Under møte for arbeidspakkeledere ved «Health Registries for Research» (HRR) 25. november 2016

4 Referanser

Databehandleravtalen mellom Helsedirektoratet og Folkehelseinstituttet signert 15.4.2016 (lagret i prosjektdokumentasjonen på F:\Forskningsprosjekter\PDB 2010 - KPR analyseprosjekte_\Prosjektdokumentasjon\Databehandleravtale)

Rammeavtalen mellom Folkehelseinstituttet og TSD ved UiO signert 22.5.2014, arkivnummer 12/632.

Avtalen mellom Folkehelseinstituttet og TSD vedrørende KPR-analyse forskningsmiljøet, signert 4.5.2016, arkivnummer 16/11181.

Rapport fra KPR-analyseprosjektet levert Helsedirektoratet 5.12.2016, korrigert versjon (lagret i F:\Forskningsprosjekter\PDB 2010 - KPR analyseprosjekte_\Prosjektdokumentasjon\Rapporter)