Perined Record linking V2.28 design notes

Samenvatting

Record linking ten behoeve van PRNinsight, gerealiseerd in Delphi (Pascal) met de volgende karakteristieken:

- Koppeling van alle bestanden (LVRh, LVR1, LVR2 en LNR) van een jaar. Het resultaat is een lijst waarin staat welke record nummers uit de bronbestanden bij elkaar horen. Alle records blijven behouden. Een gelinkt cluster kan een willekeurig aantal records uit de bronbestanden omvatten.
- De koppeling is probabilistisch. Er is een stevige statistische basis voor de beslissing of twee records dezelfde casus beschrijven of toevallig veel gelijkenis vertonen. De drempelwaarde voor zo'n beslissing is afhankelijk van de situatie en statistisch onderbouwd.
- De dichtheidsfunctie *u(w)* van een variabele, van belang voor de kans dat zijn waarde bij toeval overeenkomt met een ander, is afhankelijk van zijn waarde w. (Een moeder van 50 is onwaarschijnlijker dan een moeder van 30.)
- Verschillen tussen variabelen van gelinkte records worden beschouwd op drie niveaus: een waarde kan onbetrouwbaar zijn, een verschil kan verklaarbaar zijn (dichtbij of verwisseling van cijfers) of er is sprake van een (onverklaarbare) fout. Automatische berekening van de foutkansen.
- Grote aandacht voor de koppeling van meerlingen met inachtname van alle relevante verschillen tussen de kinderen om te voorkomen dat verwisseling van kinderen binnen een meerling optreedt.
- Alle records van alle databestanden worden op efficiënte wijze in één pass bijeen gezocht.
- Snel. Inlezen, koppelen, analyseren en rapporteren gebeurt in circa 20 sec.
- Getest. De resultaten voor 2009 en 2012 zijn vergeleken met die van de traditionele PRN koppeling.

Contents

Principes van probabilistisch koppelen

Een bronbestand bevat alle records van een dataset (LVRh, LVR1, LVR2 of LNR). Een record in zo'n bronbestand beschrijft deel-zorg van een casus. Een zwangere komt bijvoorbeeld voor in de LVR1 omdat ze tijdens haar zwangerschap is begeleid door de 1e lijn. Als ze gedurende die periode is verhuisd dan is er door haar nieuwe begeleiders opnieuw een LVR1 record aangemaakt. Is ze overgedragen naar de 2e lijn en daar bevallen, dan zal er een LVR2 record zijn dat de bevalling beschrijft. Ingeval van een meerling zelfs meerdere records. Het kind kan zijn opgenomen. Daar moet een LNR record over bestaan – of meerdere bij heropnames en overplaatsingen. Soms wordt de kraambegeleiding ook nog apart geregistreerd, weer in de LVR1. Verder kunnen er overplaatsingen zijn binnen de 2e lijn en 1e lijns zorg kan zijn verdeeld tussen huisartsen (LVRh) en verloskundigen (LVR1). Er zijn met andere woorden veel combinaties mogelijk.

Deze records hebben over het algemeen geen unieke identificatie over moeder of kind. Daarom is het bij elkaar zoeken van de records die een casus beschrijven op basis van overeenkomst van gegevens: datums, postcode, gewicht, etc. Maar deze gegevens ontbreken soms (in de LNR staat bijv. weinig over de bevalling), er worden fouten gemaakt bij het registreren (zoals een verwisseling van cijfers: 3-7-1981 wordt 7-3-1981) en de perceptie van de zorgverleners is niet altijd eender (een kinderarts scoort Apgar anders dan een gynaecoloog). Daarom betekent een verschil in de waarde van een variabele tussen twee records niet dat die records niet bij elkaar horen. Naarmate er meer overeenkomst is, is het waarschijnlijker dat ze wel dezelfde casus beschrijven. Overeenkomsten kunnen anderzijds het gevolg zijn van toevalligheden: de kans dat twee kinderen hetzelfde geslacht hebben is 50%; de kans dat ze op dezelfde dag zijn geboren is 1:365. Er bestaat een statistische grondslag om deze kansen op fouten en toevalligheden te verwerken tot een beslissing: beschrijven twee records dezelfde casus of niet. (Eigenlijk: is het waarschijnlijker dat ze dezelfde casus beschrijven dan dat ze behoren bij verschillende casus.)

Twee grootheden spelen daarbij een rol:

1. de kans dat twee waarden door toeval overeenkomen (waardoor de kans op foute links toeneemt). Deze is gelijk aan de dichtheid *u(w)* van de waarde w. Voor sommige variabelen is *u(w)* onafhankelijk van w. Bij geboortedag kind bijvoorbeeld is *u(w)* = 1/365 voor alle dagen w. Maar bij geboortedatum moeder is *u* afhankelijk van het jaar want een moeder van 50 komt minder vaak voor dan een moeder van 30.
2. De kans dat twee waarden die hetzelfde gegeven beschrijven van dezelfde casus niet overeenkomen (waardoor de kans op gemiste links toeneemt). Er is dan sprake van een fout met foutkans *e*. Deze fout kan verklaarbaar zijn volgens een bepaald recept. In datums bijv. komen verwisselingen voor van dag en maand (recept 1 met kans e_1) en soms wordt het jaar verschreven (1976 wordt 1967; recept 2 met kans e_2). Een verschil dat niet verklaarbaar is moet worden beschouwd als een onverklaarbare fout met een kans e_0.

Deze dichtheden en foutkansen worden berekend uit de resultaten van de koppeling zelf. Zie later. Ze zijn van cruciaal belang bij de beoordeling of twee records dezelfde casus beschrijven of niet. De basis voor zo'n beslissing is als volgt.

Beschouwen we twee records R1 en R2 (uit dezelfde of verschillende datasets) waarin voor een bepaald gegeven de waarden Y1 resp. Y2 is genoteerd. Er zijn drie situaties mogelijk:

Y1 = Y2	De waarden komen overeen (bijv. 3-6-1976)
Y1 ~$_k$ Y2	De waarden komen niet overeen maar het verschil is verklaarbaar volgens recept k. (Bijv. k=1: 3-6-1976 vs. 6-3-1976; k=2: 3-6-1976 vs. 3-6-1967). In het algemeen n_t recepten om verschillen te verklaren, ieder met hun eigen waarschijnlijkheid e_k.
Y1 ≠ Y2	De waarden komen niet overeen en het verschil is niet verklaarbaar. Als R1 en R2 bij elkaar horen dan is sprake van een fout, met een kans e_0.

Records R1 en R2 worden nu vergeleken op basis van twee hypothesen:

A Ze beschrijven dezelfde casus: R1 = R2
Y1 en Y2 horen overeen te komen, behoudens optredende fouten
De kans dat Y2 overeenkomt met Y1 onder de aanname dat de records dezelfde casus beschrijven is als volgt:
P(Y1 = Y2 | R1=R2) ≈ $1 - 2e_0 - 2\sum e_k$ (sommerend over alle n_t recepten)
P(Y1 ~$_k$ Y2 | R1=R2) = $2e_k$ voor alle k recepten
P(Y1 ≠ Y2 | R1=R2) = $2e_0$

B Ze beschrijven verschillende casus: R1 ≠ R2
Y1 en Y2 zijn mogelijk toevallig eender
De kans dat Y2 overeenkomt met Y1 onder de aanname dat de records niet gelieerd zijn is dan:
P(Y1 = Y2 | R1≠R2; Y1 gegeven) = $u(Y1)$
P(Y1 ~$_k$ Y2 | R1≠R2; Y1 gegeven) = $u(Y2) = u_k$
P(Y1 ≠ Y2 | R1≠R2; Y1 gegeven) = $1-u(Y1)-\sum u_k$

Deze vergelijking wordt gedaan voor alle N variabelen die in aanmerking komen. Het resultaat van die vergelijking kan worden gepresenteerd als een vector Q = (q_1,q_2,...q_N) met q_i=1 als de waarden voor variabele i overeenkomen, q_i=t als de waarden verklaarbaar verschillen volgens recept t en q_i=0 als de waarden voor variabele i niet overeenkomen en er geen recept is om het verschil te verklaren. (Variabelen waarvoor de waarde onbekend is krijgen geen element in Q.) De kans op de vector Q is afhankelijk van de hypothese:

P(Q | R1=R2) = ∏ { P(Y1=Y2|R1=R2) als q_i =1; P(Y1~$_k$ Y2|R1=R2) als q_i =t; P(Y1≠Y2|R1=R2) als q_i =0}

P(Q | R1≠R2) = ∏ { P(Y1=Y2|R1≠R2) als q_i =1; P(Y1~$_k$ Y2|R1≠R2) als q_i =t; P(Y1≠Y2|R1≠R2) als q_i =0}

Deze twee kansen zijn derhalve te berekenen voor iedere combinatie van records R1 en R2. Als P(Q | R1=R2) veel groter of veel kleiner is dan P(Q | R1≠R2) dan is de keus evident: de records horen wel of niet bij elkaar. Maar als de verschillen niet erg groot zijn dan is een nauwkeurige afweging vereist. Wat moet worden bepaald is of, bij gemeten vector Q, het waarschijnlijker is dat de records bij elkaar horen of niet. Met andere woorden: is P(R1=R2 | Q) > P(R1≠R2 | Q) ?

Die kansen kunnen worden afgeleid met het theorema van Bayes:

$$P(R1=R2 \mid Q) = \frac{P(Q \mid R1=R2) \cdot P(R1=R2)}{P(Q|R1=R2)\cdot P(R1=R2) + P(Q|R1\neq R2)\cdot(1-P(R1=R2))}$$ en P(R1≠R2 | Q) = 1 - P(R1=R2 | Q)

Met de notatie D(R1,R2) = P(Q | R1=R2) / P(Q | R1≠R2) reduceert dit tot

P(R1=R2 | Q) = D(R1,R2) · P(R1=R2) / [D(R1,R2) · P(R1=R2) +1-P(R1=R2)].

Deze kans moet groter zijn dan zijn tegengestelde, ofwel groter dan ½. Dus twee records R1 en R2 horen waarschijnlijk bij elkaar als

D(R1,R2) > (1-P(R1=R2))/P(R1=R2) ; het rechterlid noemen we de *Drempelwaarde*.

De factoren D(R1,R2) kunnen worden geschreven als het product van factoren d_i waarbij d_i voor variabele i waarden aanneemt op basis van de overeenkomst tussen Y1 en Y2 voor die variabele:

$d_i = (1-2e_0-2\sum(e_k))/u$ voor q_i=1 (overeenkomende waarden);
$d_i = 2e_k/u_k$ voor q_i=t (waarden verschillend maar verklaarbaar volgens recept k);
$d_i = 2e_0/(1-u-\sum u_k)$ voor q_i=0 (waarden onverklaarbaar verschillend).

De kans P(R1=R2) in de drempel definitie is de a-priori kans dat record R1 dezelfde casus beschrijft als R2. Deze kans is omgekeerd evenredig met de omvang M1·M2 van de datasets S1 en S2 waaruit R1 en R2 zijn getrokken en kan worden geschreven als P(R1=R2) = *alfa*(R1,R2)·M12/(M1·M2). M12 is (de schatting van) het aantal record matches, te verifiëren aan de uitkomsten van de koppeling. *Alfa*(R1,R2) duidt de waarschijnlijkheid aan dat er een record R1 in dataset S1 zit die de casus van R2 beschrijft zowel als een record R2 in S2 die de casus van R1 beschrijft. *Alfa*(R1,R2) = *alfa'*(R1,S2) · *alfa'*(R2,S1). Bijvoorbeeld:

- (R1 in LVR1, R2 in LVR2). Situatie: Kind is geboren in de 1e lijn; geen overdracht. De LVR2 zal de casus alleen beschrijven indien de moeder post partum is overgedragen terwijl dat niet in de LVR1 werd geregistreerd. Kans op postpartum overdracht =0,05; kans op niet-registratie van overdracht =0,01. *Alfa* = 0,0005.
- (R1 in LVR1, R2 in LVR2). Situatie: Overdracht naar 2e lijn. *Alfa* = 1 (LVR2 dekkingsgraad).
- (R1 in LVR1, R2 in LVR1). Situatie: Zorg eindigde om niet-medische redenen (verhuizing). *Alfa* = 1 (verwacht de zwangere bij een andere praktijk).
- (R1 in LVR1, R2 in LVR2). Situatie: Abortus. Onwaarschijnlijk dat er een LVR2 record zal zijn. *Alfa* = 0.001.

De gebruikte *alfa's* zijn in eerste instantie geschat en vervolgens geverifieerd en bijgesteld op basis van telling aan gekoppelde bestanden.

Omdat P(R1=R2) << 1, kan als besliscriterium worden gehanteerd:

D(R1,R2) > M1 · M2 / (*alfa*(R1,R2) · M12)

Dit sluit goed aan bij de intuïtie dat hogere eisen worden gesteld aan de overeenkomst tussen twee records als a) de datasets groter zijn en b) de kans kleiner is dat de casus in beide datasets is geregistreerd.

In praktijk wordt gewerkt met de logaritme van de tellers en noemers in de formule voor d_i. Dit reduceert de berekening tot een som van termen i.p.v. een product van factoren. De drempel is in dat geval *ln*(M1)+*ln*(M2)-*ln*(*alfa*(R1,R2))-*ln*(M12).

One-pass matching

Het principe is als volgt: alle records van alle datasets worden bijeengebracht in een lijst. Iedere entry in de lijst staat voor een record of voor een cluster gekoppelde records. (Initieel dus losse records, circa 350.000.) Iedere entry wordt vergeleken met alle andere entries. De twee die het meest overeenkomen (d.w.z. D(R1,R2)/Drempel(R1,R2) zo groot mogelijk) worden gekoppeld. De lijst wordt aangepast (één entry verdwijnt en het andere wordt uitgebreid met de records van de andere entry) en de vergelijking gebeurt opnieuw. Als er geen twee entries zijn die hoger scoren dan de drempel die bij hen hoort, dan is het proces gereed.

Om dit efficiënt te kunnen doen is een truc nodig: beperk de vergelijking tot entries die een kans maken om gekoppeld te worden. Ofwel: bij de behandeling van entry R1, sluit alle entries R2 uit die überhaupt *geen* kans maken om een score D(R1,R2) te behalen die groter is dan een (conservatief hoge) drempel.

Om een dergelijke strategie te kunnen volgen wordt uitgegaan van de volgende stelling: **"Als twee records dezelfde casus beschrijven zullen tenminste twee van de volgende key variabelen overeenkomen: DDGEBM, PC, DDAT, DDGEB, GEW."** Met dit uitgangspunt wordt een 'magazijn' aangelegd met negen matrices voor de mogelijke combinaties van de key variabelen. De eerste matrix bevat bijvoorbeeld waarden voor DDGEBM en PC op rijen en kolommen; de tweede DDGEBM vs. DDAT, etc. Alleen DDGEB x DDAT wordt niet gebruikt. (Een verwijzing naar) ieder record wordt opgeslagen in alle negen matrices. De stelling houdt in dat koppelbare records op tenminste één van deze matrices een identieke rij/kolom positie innemen.

Voorbeeld (kandidaten voor koppeling met R1):

record	*DDGEBM*	*PC*	*DDAT*	*DDGEB*	*GEW*
R1	6-5-1979	3312	21-9	18-9	3950
R2	5-6-1979	3312	21-9	?	?
R3	6-5-1979	?	?	18-9	3950
R4	8-10-1980	3312	8-11	8-11	3000

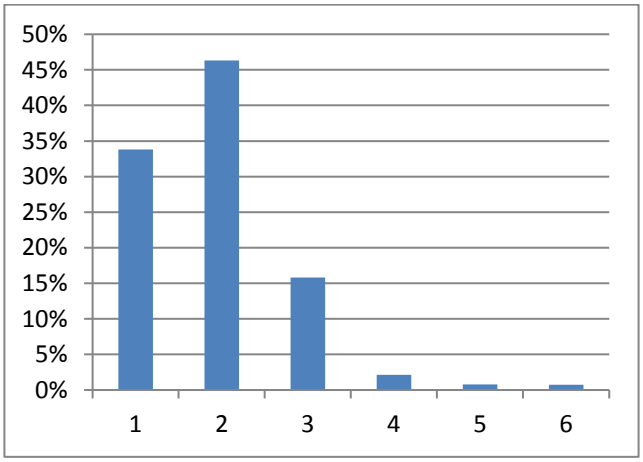
R2 en R3 zijn kandidaten voor koppeling met R1. R1 en R2 komen in dezelfde cel voor van de matrix PC x DDAT. R3 deelt een cel met R1 in DDGEBM x DDGEB. R2 en R3 komen daarom in aanmerking voor verdere analyse. Óf ze gekoppeld worden met R1 is afhankelijk van de score die vergelijking – met ook andere variabelen – oplevert en de drempel die voor hun combinatie geldt.

R4 hoeft niet te worden vergeleken met R1 want in geen enkele matrix deelt het een cel met R1.

Dus om bij gegeven record R1 een selectie te maken van kandidaat records die in aanmerking komen voor koppeling is het voldoende om de records te kiezen die in een matrix dezelfde rij/kolom positie innemen als R1 en dat te herhalen voor alle matrices. De dubbelen kunnen er uit en onderling reeds gekoppelde records worden behandeld als cluster. In praktijk leidt dat tot een beperkt aantal kandidaten. (De matrix cellen verwijzen veelal maar naar enkele records. Het hoogste aantal is rond 20.)

In principe is het mogelijk om alle bestanden in zijn geheel te koppelen door steeds een record R1 uit de lijst te nemen, alle kandidaten R2 te zoeken, de link sterktes D(R1,R2) te berekenen voor alle kandidaten en de sterkste te kiezen voor koppeling (mits hoger dan de drempel) . Maar het resultaat is (enigszins) afhankelijk van de volgorde waarin de records R1 worden gekozen. (Dit heeft te maken met de strijdigheden die kunnen voorkomen als drie of meer records worden gelinkt: A-B sterk, B-C sterk, C-A zwak.)

Een tweede selectiefase lost dit volgorde probleem op. Beschouw alle kandidaat records als een groep. Laat de onwaarschijnlijken (D<drempel/1000) hieruit weg. Alle leden van de groep kennen zelf ook weer hun eigen kandidaten. Deels komen die overeen met de reeds bekende groepsleden en deels zullen er nieuwe records bij zijn. De nieuwe records die niet onwaarschijnlijk zijn worden toegevoegd aan de groep en ook de kandidaten van de nieuw gevonden leden worden weer aan de groep toegevoegd. Dit gaat door tot alle kandidaten van alle leden van de groep in de groep zitten. Leden van deze groep kunnen derhalve niet worden gekoppeld aan records die niet in de groep zitten. Op deze wijze is de oorspronkelijke matrix van 350.000 x 350.000 record combinaties teruggebracht tot een verzameling kleinere combinaties waartussen geen verbindingen kunnen bestaan. In praktijk zijn veel van deze groepen zeer beperkt in omvang, met een enkele uitschieter tot een paar dozijn records. De figuur geeft de frequentieverdeling van groepsgroottes weer.



Verdeling van groepsgroottes

De volgorde waarin de groepen worden behandeld is onbelangrijk want er is geen interactie tussen hen.

De koppeling van records binnen een groep is op basis van ‘meerlingen eerst’, ‘sterkste link eerst’.

Voorbeeld: voor een groep van 4 records. De matrix geeft aan hoe (de logarithme van) hun onderlinge score zich verhoudt tot hun onderlinge drempel. >1 betekent score groter dan drempel; <1 betekent dat de score onvoldoende is.

	R1	R2	R3	R4
R1		**4,1**	0,7	2,0
R2	**4,1**		0,4	2,6
R3	0,7	0,4		0,8
R4	2,0	2,6	0,8	

hoogste score voor R1 x R2

	R1+2	R3	R4
R1+2		0,5	**2,5**
R3	0,5		0,8
R4	**2,5**	0,8	

Na koppeling van R1xR2

	R1+2+4	R3
R1+2+4		0,6
R3	0,6	

Na koppeling van R1xR2xR4

Het proces eindigt bij een cluster van R1+R2+R4 en een los record R3 omdat de score tussen hen onder de drempel blijft.

Groepen van slechts één record zijn al bij voorbaat gereed. Het zijn records die op geen enkele wijze een relatie kunnen hebben met andere records.

Meerlingen

Bij meerlingen is er een complicatie: de meest discriminerende variabelen hebben betrekking op de moeder (DDGEBM, PC, DDAT, KLIN) en deze is eender voor meerlingen. Onderscheid tussen kinderen kan worden gebaseerd op GEW en GEBTIJD en (in minder discriminerende mate) APGAR, SEX, LIGGING, MORT en MC. GEBTIJD en LIGGING staan niet in LNR records. MC (=meerling volgnummer) is bron van verwarring omdat kinderartsen de volgorde van geboorte niet altijd kennen en een NICU soms slechts één van de kinderen opneemt, in welk geval een volgorde nummer irrelevant is.

De laatste jaren zijn aan veel LNR records het BSN van de moeder toegevoegd alsmede de (roep)naam van het kind. Het BSN helpt om de (vele) situaties te duiden waarin OMV onbekend is of 1 terwijl er twee records van dezelfde moeder in dezelfde praktijk voorkomen. Samen met verschillen in roepnaam duidt dat dan op een tweeling. De verschillen in roepnaam zijn uitermate belangrijk om kinderen binnen een meerling van elkaar te onderscheiden. Het programma heeft uitgebreide regels om de verbastering van namen te begrijpen. Zie volgende.

Namen vergelijken

Het probleem met namen is dat er voor de spelling weinig regels zijn (JENTHE - JENTE, ROOS - ROSE), in het bijzonder bij buitenlandse namen (IBRAJIMOVIC - IBRAHIMOVICK). De verschillen die dit op kan leveren betekenen dat een overeenkomstige naam veel meer zegt over gelijkheid van het kind dan dat niet overeenkomende namen duiden op verschillende kinderen. Bij de familienamen treedt nog een extra complicatie op: de voorzetsels kunnen op diverse manieren worden weergegeven (V.D. MEER – VAN DER MEER – VAN DE MEER – MEER, V.D. – etc.).

Het programma hasht namen tot een integer op basis van fonetiek. Dit is afgeleid van de (Engelstalige) SOUNDEX met wat aanpassingen, als volgt:

1. De uitgang: ‘JE’ vervalt; ’NK’ vervangen door ‘NG’; ‘LEIGH’ vervangen door ‘LEE’.
2. Vervang dubbel-letters ‘CH’ -> ‘G’, ‘IJ’ -> ‘Y’, ‘NG’ -> ‘N’.
3. Vervang letters door cijfers:
   BFPVW -> 1
   CGJKQSXZ -> 2
   DT -> 3
   L -> 4
   MN -> 5
   R -> 6
   H -> 7 indien tussen klinkers, anders vervalt hij
   alle klinkers, non-ASCII en non-letters -> 0
4. Alle dubbele cijfers vervallen
5. Een 0 aan het eind vervalt

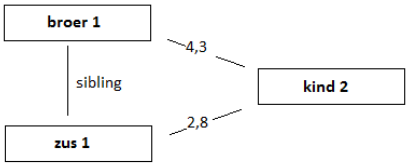
Bij de familienamen vervallen alle voorzetsels. Van alle dubbele namen wordt alleen de eerste gehashed.

Meerling koppels

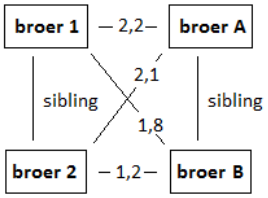
Gelukkig is er een goede registratie van meerlingen bij geboorte (doorgaans in de 2e lijn). Deze wordt gebruikt om 'sibling' links te leggen tussen records – een stap die vooraf gaat aan de overige koppelingen. Meerlingen zitten altijd in dezelfde kandidaten groep.

Als records uit een kandidatengroep worden gekoppeld kunnen vier situaties optreden:

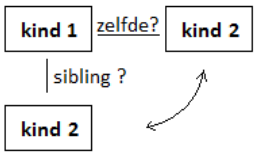
1. Geen van beide records behoort tot een meerling. Geen complicaties.
2. Eén van beide records behoort tot een meerling en heeft een sibling, de ander niet. Koppel het kind met de sterkste link. (In het voorbeeld: kind 2 met broer 1.)



3. Beide records behoren tot een meerling en hebben sibling links. Beschouw nu ook de sterkte van de links tussen hun broers/zusters. Om daarin een keus te maken worden de onderlinge link sterktes berekend voor alle mogelijke permutaties van de kinderen onderling en de keus valt op de permutatie met de hoogste som van de kind-kind link sterktes. De koppelstap resulteert in twee (of drie) koppelingen in één slag. (In het voorbeeld: hoewel 1-A de sterkste link heeft is het verstandiger om A aan 2 te koppelen en B aan 1 omdat de totaal score – 3,9 – dan hoger is dan de score van 3,4 bij A1+B2.)



4. De situatie is onhelder. OMV geeft aan dat sprake is van een meerling maar er zijn geen sibling links. (D.w.z. dat de meerling niet is gedetecteerd bij het doorzoeken van LVR2.) Is hier sprake van hetzelfde kind of zijn de twee records voor verschillende kinderen binnen een meerling? De keus wordt gemaakt op grond van overeenstemming van GEW, SEX en APGAR.



Om de complicaties tot een minimum te beperken worden meerlingen als eerste gekoppeld binnen een kandidaten groep.

Overeenkomst tussen clusters

De overeenkomst tussen twee records R1 en R2 (de sterkte van hun link) wordt uitgedrukt als D(R1,R2). Maar als een record eenmaal is gelinkt tot een cluster R1+R2 dan is de berekening voor de link met een derde record R3 – D(R1+R2,R3) – niet triviaal. En het wordt nog gecompliceerder als beide zijden van de vergelijking uit clusters bestaan, zoals D(R1+R2,R3+R4+R5).

Een simplistische oplossing zou zijn om alle onderlinge links uit te rekenen en daarvan het gemiddelde te nemen, of de laagste of hoogste of een gewogen combinatie. Maar dat gaat voorbij aan het feit dat er ook informatie zit in het gegeven dat records reeds gekoppeld zijn. Zoals:

a) Als de waarde voor een variabele in R3 onbekend is dan zal deze niet kunnen bijdragen aan een score in de vergelijking met R1 en R2. Maar als R4 en/of R5 wel een waarde hebben voor deze variabele dan hoeft hij niet als onbekend te worden beschouwd. De cluster beschrijft immers één casus dus de waarden van R4 of R5 zijn waarschijnlijk de juiste.

b) Als de waarde voor een variabele in R3, R4 en R5 identiek is en de records zijn van verschillende praktijken, dan is het onwaarschijnlijk dat die waarde toevallig is of dat er een fout is gemaakt.

Tot V19 werd de volgende strategie gehanteerd: *Bij de berekening van de link sterkte tussen clusters wordt uitgegaan van alle mogelijke combinaties van de waarden voor variabelen binnen een cluster (behalve de waarde 'onbekend' indien er ook bekende waarden zijn). De hoogste wordt gekozen.*
Maar dit heeft een bezwaar. Bij een 'zwak' cluster - met een aantal dubieuze waarden en verschillen tussen de casus in het cluster – zal het shoppen in alle mogelijke combinaties van waarden meer kans opleveren dat een match wordt gevonden. Met het risico dat het zwakke cluster wordt uitgebreid met nog meer zwakke links.
Daarom is de strategie vanaf V20 anders: de sterkte van links wordt uitgerekend tussen alle casus van het ene cluster en alle casus van het andere cluster. Daarbij worden onbekend waarden vervangen door bekende waarden uit de andere casus binnen hetzelfde cluster, mits zo'n substituut waarde uniek is (alle records hebben dezelfde waarde of onbekend). Van alle combinatie scores wordt het gemiddelde genomen.

Indien een waarde voor DDGEBM, PC of GEW uniek is binnen een cluster en meer dan eens voorkomt en evenzo binnen het andere cluster, dan wordt bij onderling verschil een korting opgelegd.

Voorbeeld: Records R1, R2 en R3 zijn gekoppeld tot een cluster en evenzo R4 en R5.

record	*DDGEBM*	*PC*	*DDAT*	*DDGEB*	*GEW*
R1	6-5-1979	3312	21-9	18-9	3950
R2	5-6-1979	3312	21-9	?	?
R3	6-5-1979	?	?	18-9	3900
R4	8-10-1980	3311	8-11	8-11	3900
R5	8-10-1980	3311	9-11	9-11	?

Als R4 of R5 gematched wordt met R2 dan zal voor R2 DDGEB de waarde 18-9 worden aangenomen maar de waarde voor GEW blijft onbekend. Als R5 wordt gematched met R3 dan zal voor R3 worden aangenomen dat PC=3312 en DDAT=21-9; R5:GEW zal op 3900 worden gesteld.

Dubieuze gegevens

Drie categorieën worden gedefinieerd voor de status van een waarde voor een variabele:

a) De waarde is onbekend
b) De waarde is bekend maar dubieus
c) De waarde is bekend en realistisch

Categorie b, dubieus, speelt een rol bij de berekening van de link sterkte D(R1,R2). Beschouw het voorbeeld van DDGEBM voor een zwangere die bij haar eerste bevalling 52 jaar zou zijn geweest. Dat is mogelijk maar niet erg waarschijnlijk. Als er een koppeling is met een ander record waarin dezelfde

geboortedatum wordt gemeld dan is de waarde waarschijnlijk wel correct. Maar als het andere record meldt dat ze 32 was dan is er waarschijnlijk sprake van een fout. Die hoeft niet verklaarbaar te zijn volgens een van de recepten die zijn geïmplementeerd om verschrijvingen te detecteren.

Maar indien DDGEBM een leeftijd zou aangeven van 72, dan zou overduidelijk sprake zijn van een fout en de waarde kan dan worden vervangen door 'onbekend'. Een onbekende waarde heeft geen invloed op de link sterkte maar een mismatch wel.

Door een waarde w de status 'dubieus' te geven kan een middenweg worden bewandeld:

- Bij een match met de waarde van het te koppelen record wordt w beschouwd als correct en draagt zijn match bij aan D(R1,R2) via $q_i = 1$ in de vector Q;
- Bij mismatch wordt aangenomen dat de waarde onbekend is. Hij draagt dan in het geheel niet bij aan Q.

De criteria voor deze dubieuze status zijn als volgt:

Variabele	Criterium voor 'dubieuze' status	Incidentie
DDGEBM	De geboortedatum van de moeder is dubieus indien zij bij geboorte 50 jaar of ouder is of jonger dan 12+pariteit.	LVR: 0,01% LNR: 0,1%
DDGEB	De geboortedatum van het kind is dubieus indien het jaar niet overeenkomt met het jaar van registratie. In de LVR2 wordt bovendien gecheckt of DDGEB en het partusnummer in harmonie zijn. Omdat in de LVR2 wordt geregistreerd bij geboorte betekent een hoger partusnummer over het algemeen ook een latere geboortedatum - behoudens 'breuken' in het partusboek t.g.v. aparte series (zoals voor OK). Records worden daartoe gesorteerd op partusnummer. Indien voor een record geldt dat het voorliggende én het naliggende record een DDGEB hebben die beide hoger of beide lager zijn, dan wordt DDGEB gemarkeerd als dubieus.	LVR: 3%
DDAT	Als DDGEB bekend is dan wordt DDAT als dubieus beschouwd indien DDGEB-DDAT groter is dan 30 of kleiner dan -75 bij een gewicht>1500. Als DDGEB onbekend is dan wordt dezelfde test toegepast met DDVLIES-DDAT.	LVR: 0,03% LNR: 0,1%
PAR	Een graviditeit of pariteit >10 wordt als dubieus gemarkeerd.	LVR: 0,2%
PC	Er zijn twee criteria om een postcode als dubieus te beschouwen: 1) de postcode bestaat niet of heeft geen adressen en 2) de afstand van de zorgverlener tot de postcode is groot. Om dit te testen wordt gebruik gemaakt van een postcodelijst met de coördinaten van het midden van het gebied. Voor een zorgverlener wordt een centraal punt uitgerekend op basis van de coördinaten van alle records voor die zorgverlener. In de 1e lijn hanteren we een grens van 25 km tot dit centrale punt (afgeleid uit de LVR1 zelf). Als een postcode daarbuiten ligt en er zijn minder dan 5 records voor die postcode bij deze praktijk en er zijn tenminste drie andere praktijken die dichterbij liggen, dan wordt de postcode als dubieus gemarkeerd. Voor de LVR2 doen we hetzelfde met een grens van 70 km en voor LNR met een grens van 150 km. Voor NICU overplaatsingen is dit criterium discutabel. Maar het schaadt de koppeling niet.	LVR: 0,4% LNR: 0,1%

KLIN	LVR2 praktijknummers spelen een rol in de koppeling: bij verwijzing vanuit LVRh en LVR1 wordt de kliniek vermeld waarnaar werd overgedragen en in de LNR wordt het nummer vermeld van de kliniek waar de partus plaatvond. In een aantal gevallen zijn deze praktijknummers dubieus. De volgende situaties doen zich voor:		
	Nummer ongeldig	Het opgegeven praktijknummer kan niet behoren bij een LVR2 praktijk omdat het valt buiten de range 1-999. Of het opgegeven nummer zou van een LVR2 praktijk kunnen zijn, maar die praktijk bestaat niet en het nummer komt ook niet voor in een lijst van vervallen nummers vanwege fusies.	LVR1: 1% LVR2: 1% LNR: 5%
	Fusie	Het opgegeven nummer is van een praktijk die is opgegaan in een andere praktijk (fusie). Over het algemeen is de transformatie dan eenduidig en kan het praktijknummer worden gewijzigd. In een aantal gevallen echter wordt binnen een gefuseerde praktijk nog steeds onderscheid gemaakt tussen lokaties. Deze situaties krijgen een speciale behandeling bij optredende mismatches: als verandering van het praktijknummer in dat van de fusiepartner de mismatch opheft, dan wordt dat behandeld als match met een kleine straf.	
	Geen records	De LVR2 praktijk heeft niet geregistreerd. We kunnen dit slechts vaststellen - betreffende records zullen niet gekoppeld kunnen worden aan de LVR2.	

Verklaarbare fouten

Zoals besproken bij Principes van probabilistische koppeling kan de vergelijking tussen de waarden van een variabele in twee records leiden tot drie situaties:

a) De waarden komen overeen.
b) De waarden zijn verschillend maar er is een recept om het verschil te verklaren (een vorm van systematische fout).
c) De waarden zijn verschillend en geen enkele recept kan het verschil verklaren; dit zijn toevallige fouten.

Toevallige fouten zijn onverklaarbaar. Waarom staat in record R1 dat het om een meisje gaat terwijl record R2 een jongen aangeeft? Hoe kan het dat het ene record moeders geboortedatum stelt op 3-5-1978 en het andere op 18-4-1980? Deze fouten zijn wel reëel. Ze komen voor in paren van records die in alle andere opzichten identieke gegevens bevatten. De meest voordehand liggende verklaring is dat de informatie is overgenomen van een andere bron en dat daarbij de bron is verwisseld. Hoe dan ook: onverklaarbare fouten komen voor en de probabilistische analyse bestraft ze met een negatieve bijdrage voor de link sterkte D(R1,R2).

Verklaarbare fouten zijn van een ander karakter. Ze zijn verklaarbaar onder de aanname van een ‘recept’. Een recept is gegrond op intuïtie en kan worden onderbouwd met metingen. De recepten die worden gehanteerd staan in de tabel hieronder.

Het effect van een verklaarbare fout op de link sterkte is eveneens een straf, in de vorm van een negatieve bijdrage aan D(R1,R2). Maar omdat de waarden 'na correctie' overeenkomen ontvangt D ook de positieve bijdrage van overeenkomende variabelen. In praktijk betekent dit veelal dat een verklaarbaar verschil nauwelijks invloed heeft op D (net zoals een onbekend gegeven geen invloed heeft).

Voorbeeld:

Twee DDGEBM (van 30-jarige zwangeren) die overeenkomen contribueren 8,4 aan D. Zijn ze verschillend maar komen ze overeen als dag en maand worden verwisseld, dan gaat hier een straf vanaf van 9,2. Het verklaarbare verschil leidt dus tot een bijdrage van -0,8. Zou het verschil als niet verklaarbaar worden beschouwd dan is er alleen maar straf: de bijdrage aan D is dan -8,4.

Variabele	Verklaarbare fouten
KLIN	De LVR2 praktijknummers zijn in de loop der jaren gewijzigd door fusies. Hier wordt überhaupt rekening mee gehouden: LVR1 records die verwijzen naar LVR2 praktijknummers die door fusies zijn veranderd, worden aangepast. Maar er zijn ook situaties waar binnen een fusie nog steeds twee aparte lokaties worden gecodeerd (zoals Deventer en Apeldoorn). Als verklaarbare fout wordt hier een verwijzing beschouwd naar een andere lokatie binnen dezelfde fusie. Hetzelfde geldt voor de registratie van de partus kliniek in de LNR.
DDGEB DDAT DDVLIES	Voor deze datums geldt een 'period of grace': enkele dagen verschil in de waarden van records wordt behandeld als een verklaarbare fout. De omvang van deze period of grace wordt bepaald aan de hand van meting aan de LVR1-LVR2 gekoppelde records. Met name DDAT is gevoelig voor deze behandeling. Dat komt omdat deze datum in de LVR1 kan zijn ingevuld vóór de 20w echo terwijl de LVR2 de aangepaste datum heeft die uit de echo volgde.
DDGEBM	Een aantal verschrijvingen wordt herkend en behandeld als verklaarbare fout: een verschil van 1 cijfer in dag of maand, verwisseling van dag en maand en verwisseling van de laatste twee cijfers in het geboortejaar.
GEBTIJD	Afronding van de tijd op 5 minuten en een vol uur verschil worden herkend als verklaarbaar.
GEW	Afronding naar 0, 00, 5 en 50 worden herkend als verklaarbaar.
PC	Een paar procent van de zwangeren verhuist voor de geboorte. Dit kan verklaren dat de postcode die de verloskundige optekende (bij intake) afwijkt van de postcode die in de 2e lijn werd geregistreerd (bij geboorte). Omdat het gros van de verhuizingen plaatsvind over korte afstanden herkennen we postcode verschillen als verklaarbaar indien de twee postcodes niet meer dan 7 km uit elkaar liggen.
APGAR	We accepteren een verschil van 1 punt als verklaarbaar.

Bepaling van dichtheden *u*

De dichtheden *u* kunnen worden bepaald uit de bronbestanden. Voor de meeste variabelen kunnen de waarden worden ingedeeld in groepen. Door te tellen in een bronbestand wordt vastgesteld hoe vaak een waarde in een groep valt. De tabel geeft de implementatiekeuzen aan en gevonden

waarden voor de bestanden van 2012. NB: de tabel geeft alleen de extreme waarden aan (hoogste waarde is voor minst voorkomende waarde *w*; laagste waarde voor meest voorkomende waarde).

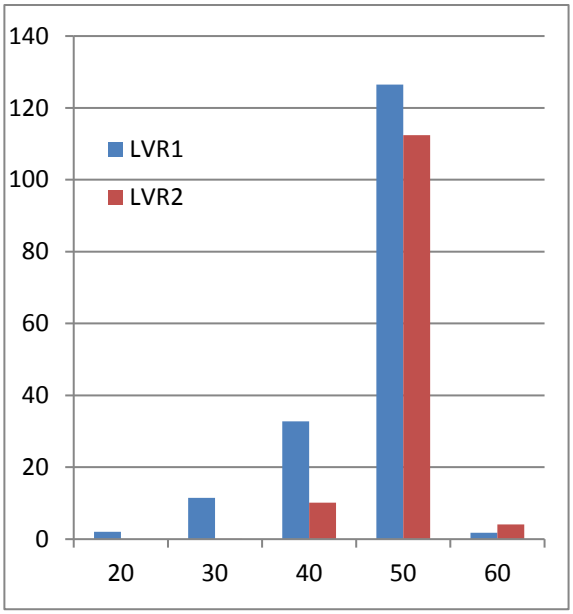
Variabele	Differentiatie	*-ln(u)*
GEW	Indeling in groepen van 100 gram. Geteld in LVR1 + LVR2.	2,2 – 9,2
	Gewicht eindigend op 0	2,7 extra
	Gewicht eindigend op 5	3,7 extra
	Gewicht eindigend op ander cijfer	6,9 extra
PC	Indeling in groepen van 100. Geteld in LVR1 + LVR2.	7,4 – 9,2
DDGEBM	Onderscheid naar jaar. Geteld in LVR1 + LVR2.	8,4 – 9,2
APGAR	Iedere waarde (1-10) wordt geteld in LVR2 + LNR.	0,4 – 6,6
PAR	Iedere waarde wordt geteld in LVR1 + LVR2.	0,7 – 8,2
HER	Onderscheid wordt gemaakt tussen Nederlands en Overig.	0,2 – 1,6
MORT	Onderscheid wordt gemaakt tussen wel/niet overleden in LVR2.	0,0 – 4,3
DDAT	Uniforme verdeling.	5,9
DDGEB	Uniforme verdeling.	5,9
DDVLIES	Uniforme verdeling.	5,9
PERIODE1	Het aantal dagen tussen DDGEB en DDAT (d.w.z. zwangerschapsduur-280) krijgt ieder zijn eigen dichtheid op basis van telling in LVR1+LVR2.	3,2 – 9,2
PERIODE2	Het aantal dagen tussen DDGEB en DDVLIES krijgt ieder zijn eigen dichtheid op basis van telling in LVR1+LVR2.	1,7 – 9,2
LIGGING	Onderscheid wordt gemaakt tussen Hoofdligging en Overig. Telling in LVR1 + LVR2.	0,0 – 3,0
SEX	Jongen/meisje.	0,7
GEBTIJD	Uniforme verdeling over het etmaal.	7,3
KLIN	Uniforme verdeling	4,6

Verder wordt rekening gehouden met de correlatie tussen gewicht en zwangerschapsduur ($u_{gew}(w)$ bestaat uit afzonderlijke tabellen voor zwangerschapsperioden <200d, 200-229d, 230-249d, 250-279d en ≥280d).

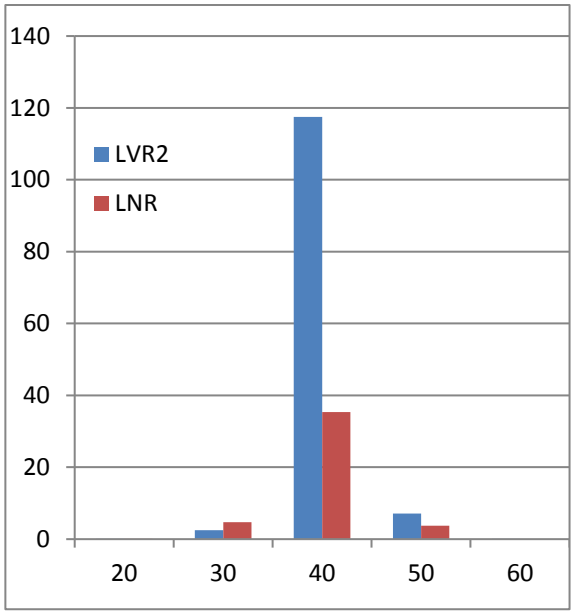
Door te tellen in LVR1 + LVR2 ontstaat een overschatting van bijzondere liggingen en van extreme gewichten. Die casus worden immers vaker overgedragen. Deze overschatting is acceptabel. Als een *u* in werkelijkheid half zo groot is dan zal *-ln(u)* 0,7 toenemen.

Voor de koppeling tussen LVRh, LVR1 en LVR2 records komen de volgende variabelen in aanmerking: DDGEBM, PC, DDAT, DDVLIES, PAR, HER, GEW, GEBTIJD, MORT, DDGEB, LIG, KLINOVRDR, SEX en de correlatieperiodes PERIOD1 en PERIOD2. Voor de koppling van een LNR record komen de volgende variabelen in aanmerking: DDGEBM, PC, DDAT, GEW, DDGEB, KLINPARTUS, APGAR, SEX. DDAT voor LNR records wordt afgeleid uit de geregistreerde zwangerschapsduur en DDGEB.

Per record kan worden bepaald welke score maximaal mogelijk is bij koppeling. Dat is namelijk de interne score die ontstaat bij koppeling met het record zelf. De figuren hieronder geven de verdeling van die maximale scores weer voor koppelingen op basis van de twee sets variabelen.



Interne score (LVR–LVR variabelen)



Interne score (LVR–LNR variabelen)

Het is uit de grafieken onmiddellijk duidelijk dat koppeling van en aan LNR records veel minder nauwkeurig zal zijn dan koppeling van LVR records onderling.

Schatting van foutkansen *e*

Als eenmaal een groot aantal records is gekoppeld dan kan worden geteld hoe vaak een fout optreedt. Maar om te kunnen koppelen moeten de foutkansen *e* bekend zijn. Dit kip-ei probleem wordt opgelost via iteratie. Daartoe nemen we in eerste instantie aan dat alle *e* nul zijn. LVR1 records worden vervolgens simplistisch gekoppeld aan LVR2 records met aanname van een (veel te hoge) drempel. Dit leidt derhalve tot record combinaties die een zeer sterke overeenkomst hebben. Sommige variabelen kunnen niettemin in waarde verschillen. Die verschillen worden toegewezen aan verklaarbare fouten en onverklaarbare fouten. Hieruit volgen de e_0 en e_t voor de diverse variabelen. Met die waarden bekend wordt de koppeling herhaald, nu met een lagere drempel. (De drempel wordt tijdens de iteratie verlaagd om stabiliteit te garanderen. In principe kan zo'n iteratie ook exploderen: grotere fouten leiden tot slechtere koppels en die suggereren weer grotere fouten, etc.) Over het algemeen leidt dit tot een wat andere koppeling waaruit weer nieuwe schattingen volgen voor de foutkansen. Na een paar slagen eindigt deze iteratie omdat de uitkomsten niet meer veranderen.

De simplistische koppeling implementeert niet de tweede fase van kandidaat selectie. Voor ieder record uit de LVR2 worden kandidaten gezocht uit de LVR1 via het 'magazijn' principe zoals eerder beschreven. Van deze kandidaten wordt de linksterkte bepaald en de hoogste wordt gekoppeld (voorzover boven de drempel).

Eenzelfde procedure wordt gevolgd voor bepaling van de foutkansen bij de variabelen die een rol spelen in LNR koppelingen. Over het algemeen zullen foutkansen bij LVR koppelingen anders zijn dan bij LNR koppelingen, ook al gaat het over dezelfde variabele.

Typische waarden voor de (logarithme van) foutkansen staan in onderstaande tabel. De ranges bij verklaarbare fouten geven de extremen aan die optreden bij de diverse recepten die zijn gedefinieerd.

Variabele	LVR koppeling		LNR koppeling	
	$-ln(e_0)$	$-ln(e_t)$	$-ln(e_0)$	$-ln(e_t)$
DDGEBM	8,7	9,2	9,2	7,7
PC	5,3	5,4	6,2	6,1
DDAT	7,4	4,2-9,2	2,4	9,2
DDVLIES	5,2	3,7-8,2		
PAR	2,5			
HER	3,0			
GEW	4,2	7,5	4,2	7,3
GEBTIJD	3,6	5,9		
MORT	7,0			
DDGEB	6,2	4,7-7,6	9,2	6,0-9,2
LIGGING	5,3			
KLINOVDR	4,1			
KLINPARTUS			5,4	
APGAR			4,7	3,7
SEX	5,7		4,5	

Merk op dat 9,2 = -*ln*(0,0001) het maximum is dat wordt gehanteerd. Als minder fouten worden geteld dan wordt toch aangenomen dat er wel een kans van 1:10.000 zal zijn op zijn optreden.

Bijzondere situaties

In enkele situaties wordt afgeweken van de bovenbeschreven strategie, namelijk:

- **LVR1:Abortus**. Van abortus casus wordt niet verwacht dat er een overeenkomstig record bestaat in de LVR2. Deze records worden daarom terzijde gelegd tot de koppeling gereed is. Daarna wordt voor alle zekerheid nog getracht om abortus records te koppelen aan LVR2 records. (Daarbij wordt geëist dat DDAT vergelijkbaar is. Anders bestaat het risico dat wordt gekoppeld aan een record van een volgende zwangerschap.)
- **LVR1:Verhuisd (vertrokken)**. Over het algemeen wordt 'Einde zorg om niet-medische reden' in de LVR1 rubriek 'Reden einde zorg' gezien als indicatie dat cliënt is vertrokken vanwege verhuizing of ontevredenheid. Zulke records bevatten behalve PC, DDGEBM en DDAT weinig informatie. Ze worden daarom terzijde gelegd tot de koppeling gereed is. Daarna wordt getracht ze te koppelen aan alle overige records.
- **LNR clusters**. Bij ieder LNR record hoort een link te worden gevonden naar LVR1/h en/of LVR2. Als de koppeling van kandidaat groepen tot clusters resulteert in een cluster met alleen LNR records dan wordt de drempel een paar punten verlaagd waarna opnieuw wordt gekeken of de cluster kan worden uitgebreid. Als alle clusters behandeld zijn wordt nogmaals gepoogd om losse LNR clusters te koppelen door veel waarde te hechten aan overeenkomend BSN van de moeder.
- **LVR1 clusters**. Eenzelfde procedure wordt gevolgd voor losse LVR1 records met de indicatie dat ante partum of durante partu overdracht naar de 2e lijn heeft plaatsgevonden of dat de

zorg beperkt is gebleven tot alleen de kraamzorg. Van deze records mag immers worden verwacht dat er een corresponderend LVR2 record (bij kraam ook LVR1) bestaat.

De namen in LNR records zijn van wezenlijk belang bij de koppeling van kinderen binnen een meerling. Maar namen kennen veel verschillende notaties. Zo wordt bijv. “VAN DEN BERG” ook wel genoteerd als “VDBERG”, “BERGVD”, V.D.BERG” en “BERG”. Veel buitenlandse namen worden soms verbasterd: “JESSIE” komt ook voor als “JESSY”en “JESSEY”. Ondanks een zeer uitgebreid algorithme om notatieverschillen te behandelen als verklaarbaar verschil, blijven er situaties die op het oog wel gemakkelijk zijn te interpreteren als identiek of verschillend. Bij de berekening van D kan alleen overeenkomst en verklaarbaar verschil tussen namen leiden tot een bijdrage. Een onverklaarbaar verschil tussen namen (zoals “ROOS” en “ROSALIE”) leidt niet tot een straf.

Loose ends

Over het algemeen zal een LVR1 bestand van het ‘vorig jaar’ (d.w.z. het jaar voorafgaand aan het koppeljaar) worden toegevoegd. Daarvan worden alleen records beschouwd die kunnen resulteren in een geboorte rond de jaargrens. Het is daardoor mogelijk dat een link wordt berekend tussen records van verschillende jaren. Maar de linkfile had vroeger betrekking op geboorten/abortus van één jaar. Alle records die gelinkt zijn aan casus met einde zwangerschap in een ander jaar waren toen niet terug te vinden in de linkfile. Ze werden door de datamanager van Perined Insight gerapporteerd in een apart bestand met de naam LooseEnds.txt.

Program flow

Lees instructiebestand (waarin staat welke bestanden, namen van variabelen, etc.)
Lees databestanden (tab-separated *.TXT of SPSS *.SAV)
Strip abortus en verhuisd
Maak data magazijnen voor eenlingen uit LVR1, LVR2 en LNR
Bereken *u’s* met gegevens uit deze datamagazijnen
Bereken *e’s* met gegevens uit deze datamagazijnen via iteratie
Maak magazijn met alle kinderen (behalve abortus en verhuisd)
Voor ieder record dat nog niet is behandeld:
    Maak kandidaten groep
    Identificeer ‘sibling’ links
    Bereken link sterkte t.o.v. drempel tussen alle records in de groep -> matrix
    Link meerlingen (steeds sterkste binding; herbereken matrix); herhaal
    Voeg eenlingen toe en link alle (idem); herhaal
    Speciale behandeling voor ongekoppelde LNR en LVR1 clusters
    Markeer alle kandidaten als behandeld
Link abortus en verhuisd
Link overgebleven losse LNR clusters op basis van BSN en naam overeenkomsten
Output
Analyse

Tests

“False positives” zijn betrekkelijk eenvoudig op te sporen door binnen gekoppelde clusters de onderlinge link sterktes te berekenen. Zwakke links moeten verklaard kunnen worden want anders duiden ze op koppelfouten.

“False negatives” zijn veel moeilijker op te sporen. Het zijn de links die ontbreken terwijl records wel bij elkaar horen. Om dit te testen zijn de uitkomsten vergeleken met uitkomsten van de traditionele PRN koppeling voor 2009 en 2012. Dit detecteert enige ontbrekende links in het ‘grijze gebied’, waar op het oog een link wel of niet zou worden gelegd. (Het signaleert een aantal systematische fouten in de PRN koppeling, zowel false positives als false negatives, zowel bij eenlingen als bij meerlingen).

Voorbeeld output

De output bestaat uit een lijst met per regel de record nummers van records die aan elkaar zijn gekoppeld. Het bronbestand wordt aangeduid met ‘h’, 1, 2 of ‘K’ gevolgd door het jaar. Iedere regel begint met een code die weergeeft hoeveel records uit de vier datasets op die regel voorkomen, in de volgorde h-1-2-K.

Circa 1/3 van de ‘links’ bestaat uit losse LVR1 records (waarvan 1/6 abortus); 40% wordt gevormd door LVR1-LVR2 koppels; 15% door LVR1-LVR2-LNR clusters; 4% door losse LVR2 records. Het grootste cluster bestaat 10 records: LVR1, LVR2 en 8x LNR. 3% van de clusters heeft meer dan één LVR1 record. Rond 600 LNR clusters konden niet worden gelinkt aan LVR1/h of LVR2.

Voorbeeld:

```
datasets (LVRh,LVR1,LVR2,LNR)+(per record): dataset(h/1,2,K)-jr-recNr

0001    K-2012-10442
1100    h-2012-113      1-2012-155239
0110    1-2012-1005     2-2012-123146
1001    h-2012-304      K-2012-32837
1010    h-2012-1        2-2012-114846
1111    h-2012-249      1-2012-181251    2-2012-23025    K-2012-15793
0221    1-2012-35214    1-2012-89570     2-2012-116669   K-2012-41776
```

Daarnaast worden nog twee bestanden aangemaakt:

- (LNR) heropnames. Dit bestand toont per regel de recordnummers van bij elkaar horende LNR records, voorafgegaan door het aantal. Voorbeeld:

```
3     9570  9632  9633
3     973   1048  29576
5     1776  1780  1775  1773  18115
5     19456 7891  7868  7887  15573
```

- Research file. Het bestand toont per casus één record uit een dataset (indien aanwezig) waarbij abortus en verhuizing wordt uitgesloten en DDGEB bekend moet zijn (als indicatie dat er een kind is). Voorbeeld:

```
LVRh-jr LVRh-recNr     LVR1-jr LVR1-recNr     LVR2-jr LVR2-recNr     LNR-jr  LNR-recNr
                       2014    117962
                       2013    7963           2014    22806          2014    15181
                       2014    117965         2014    75422          2014    226
```