Registratiegraad van in ziekenhuizen opgenomen verkeersslachtoffers

Eindrapport

R-97-15
Dr. P.H. Polak
Leidschendam, 1997
Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV

Documentbeschrijving

Rapportnummer:	R-97-15
Titel:	Registratiegraad van in ziekenhuizen opgenomen verkeersslachtoffers
Ondertitel:	Eindrapport
Auteur(s):	Dr. P.H. Polak
Onderzoeksmanager:	Drs. S. Oppe
Projectnummer SWOV:	53.214
Projectcode opdrachtgever:	BPVL 95.005
Opdrachtgever:	De inhoud van dit rapport berust op gegevens verkregen in het kader van een project, dat is uitgevoerd in opdracht van de Adviesdienst Verkeer en Vervoer van Rijkswaterstaat. Het onderzoek is uitgevoerd in samenwerking met SIG Zorginformatie, Utrecht. Onderzoek en analyse: Dr. D.H.M. Frijters Software engineering: Ir. S.A. Westen
Trefwoord(en):	Data bank, recording, accident, injury, data acquisition, data processing, calibration, hospital, first aid, statistics, quality assurance, classification, Netherlands.
Projectinhoud:	De politie registreert niet alle verkeersongevallen en het is zowel politiek als beleidsmatig gewenst het werkelijke aantal zo goed mogelijk te kennen. Tegen deze achtergrond is onderzoek verricht naar de registratiegraad van verkeersslachtoffers die in een ziekenhuis zijn opgenomen. De uit het onderzoek verkregen kennis over codeerfouten bij de registraties (de VerkeersOngevallenRegistratie VOR en de Landelijke Medische Registratie LMR), is samengevat in een fouten-catalogus. Aanbevelingen zijn gedaan ter verbetering van de kwaliteit van beide bestanden.
Aantal pagina's:	126 p. + 28 p.
Prijs:	f 45,-
Uitgave:	SWOV, Leidschendam, 1997

Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV
Postbus 170
2260 AD Leidschendam
Telefoon 070-3209323
Telefax 070-3201261

Postadres gewijzigd in:
Postbus 1090
2260 BB Leidschendam

Samenvatting

De aanleiding voor dit onderzoek naar de registratiegraad van verkeersslachtoffers die in een ziekenhuis zijn opgenomen (de doelpopulatie), is dat bekend is dat de politie niet alle verkeersongevallen registreert en dat het zowel politiek als beleidsmatig gewenst is het werkelijke aantal zo goed mogelijk te kennen.
Een van de manieren om tot een verantwoorde ophoging te komen, is koppeling van de VerkeersOngevallenRegistratie (VOR), op grond van opgaven van de politie, aan de Landelijke Medische Registratie (LMR), gebaseerd op in principe alle uit ziekenhuizen ontslagen patiënten. Hierop is het onderhavige onderzoek gericht geweest.
Een dergelijke koppeling is eerder uitgevoerd. In de veronderstelling dat er bij het registratieproces codeerfouten gemaakt kunnen worden en dat daardoor ten onrechte matches van records die dezelfde persoon betreffen niet plaatsvinden, is in dit project besloten bepaalde verschillen toe te laten bij de koppeling. Verder is, anders dan de vorige keer, het gehele VOR-bestand in het onderzoek betrokken en een grotere subset uit de LMR.

Een subset van de bestanden is beoordeeld op verschillen in coderingen van de zogenaamde koppelvariabelen, waarbij aan gevonden codeerfouten een verschillend gewicht is toegekend: hoe ongebruikelijker de fout, des te groter het gewicht. De aan fouten toegekende gewichten zijn meegenomen in een afstandsfunctie, waarmee de afstand tussen records uit enerzijds de VOR en anderzijds de LMR is bepaald.

Door nu koppeling nog toe te staan tot een bepaalde afstand, wordt de koppeling ongevoelig gemaakt voor veel voorkomende fouten. Ieder record wordt gekoppeld met zijn naaste buur uit het andere bestand, maar de afstand tussen die naaste buur en de op één na dichtstbijzijnde (de selectiviteit) wordt ook behouden. Zo wordt informatie verkregen over de uniekheid van de koppeling.
De gekoppelde records zijn ingedeeld naar mate van zekerheid dat de koppeling terecht heeft plaatsgehad. Daarbij is gebruik gemaakt van de afstand waarop gekoppeld is en de selectiviteit van deze match. Deze zekerheid liep van 100% voor met afstand 0 en grote selectiviteit gekoppelde records tot slechts enkele procenten als een grote afstand bestond.
Om vervolgens ophoogfactoren te kunnen bepalen, is een nieuwe methode ontwikkeld, de 'footprint-methode'. Daarbij is informatie over de vervoerswijze van de slachtoffers gebruikt, die bij de koppeling geen rol heeft gespeeld. Deze geeft onafhankelijke informatie over de juistheid van de koppeling.
De vervoerswijze wordt bij LMR en VOR volgens sterk verschillende codeerinstructies en -conventies geregistreerd. Dit leidt bij de met afstand 0 gekoppelde records tot een duidelijk patroon van combinaties van de twee codeerwijzen: de 'footprint'. Deze footprint is nu gebruikt om de aandelen correct gekoppelde records te bepalen in de met grotere afstand gekoppelde groepen records.

De informatie die deze analyse heeft gegeven is gebruikt om de omvang van de doelpopulatie vast te stellen in de beide bestanden. Ook is een schatting gemaakt van het deel dat in beide registraties ontbreekt. Hiermee kon de

ophoogfactor voor alle in een ziekenhuis opgenomen verkeersslachtoffers bepaald.
Vervolgens zijn ophoogfactoren berekend voor twee indelingen van de slachtoffers waarvan de registratiegraad sterk afhankelijk is, namelijk de wijze van verkeersdeelname van het slachtoffer en de provincie waar het ongeval heeft plaatsgehad.

De resultaten zijn dat, gebaseerd op de bestanden over 1992 en 1993, het aantal ziekenhuisgewonden dat de politie opgeeft (gemiddeld 12065 per jaar), met een factor 1,64 moet worden vermenigvuldigd om het werkelijke aantal (gemiddeld 19745 per jaar) zo goed mogelijk te benaderen. Dit is circa 2% meer dan het gebruikelijke aantal verkeersslachtoffers volgens de LMR.
De ophoogfactoren naar wijze van verkeersdeelname lopen uiteen van 1,3 voor auto-inzittenden tot 2,5 voor fietsers. Voor provincies variëren de ophoogfactoren van 1,3 voor Zeeland tot 2,1 voor Friesland. Deze factoren zijn geldig voor de onderzochte jaren, maar aannemelijk is dat ze niet sterk fluctueren, zodat de verkregen factoren een aantal jaren gebruikt kunnen worden. Wel dient de koppeling op gezette tijden herhaald te worden om veranderingen in de registratiegraad vast te stellen.

De uit het onderzoek verkregen kennis over codeerfouten bij de registraties is samengevat in een foutencatalogus. Aanbevelingen zijn gedaan ter verbetering van de kwaliteit van beide bestanden.

Het onderzoek heeft tevens geleid tot de conclusie dat de ontwikkelde koppelingsmethode onder bepaalde voorwaarden ook gebruikt kan worden om andere bestanden te koppelen. Dit geldt met name voor bestanden waarbij veel informatie per record beschikbaar is, maar niet alle informatie foutloos of volledig wordt geregistreerd.

Summary

The Registration Rate of Hospitalised Road Accident Victims

The reason for carrying out this research project into the registration rate by the police of those victims admitted to hospital (the target population) is that it is well known, that they do not record every accident. Both the politicians and the policy makers wish to know the actual number, as far as this is possible.
One of the ways of responsibly extrapolating the recorded numbers is to link the police data - as processed by the Traffic Accident Data Administration (TADA) of the Ministry of Transport - with the National Patient Register (NPR) of the Ministry of Health. The latter registers, in principle, áll discharged hospital patients. This project was based on this principle.
A similar linkage had already been carried out. This time certain differences between the two databases were tolerated, because it was assumed that coding mistakes occur during the processing whereby certain linkages were not made, in spite of the fact that they both concerned the identical person. Furthermore, this time the complete TADA file was used together with a larger sub-set of the NPR.

The validity of the two databases was judged using a sample set of the files and searching for coding differences of the so-called link variables. Discovered coding mistakes were then given a certain weighting: the more unusual the mistake, the bigger the weighting. The weightings of mistakes made were then used to calculate a so-called distance function. Here the distance between TADA and NPR records was calculated.

By permitting a certain maximum distance, the linkage was made unsusceptible to common mistakes. Every record was linked to its immediate neighbour from the other database. However, the distance between this immediate neighbour and the next closest (the selectivity) was also registered. Information was thus obtained over the uniqueness of the linkage. The linked records were assorted according to the extent of the certainty that the linkage was correct. The distance of the linkage and the selectivity of the linkage were used for this. This certainty varied from 100%, for linkage with a zero distance and a high selectivity, to a small percentage if the distance was great.
In order to be able to calculate the extrapolation factors, a new method was developed; the 'footprint method'. For this the patient's modal split was used, even though it had played no part in the actual linkage. This provided independent information about the validity of the linkage.
TADA and NPR use very different definitions and coding instructions to determine the modal split. This resulted, for those records linked, having a zero distance, in a clear pattern of combinations of the two ways of coding; the 'footprint'. This footprint was used this time to calculate the shares of the correctly linked records in those groups with a larger distance.

The information thus gathered was used to determine the size of the target population in both databases. An estimate was also made of that part missing in both databases. Using this, it was possible to calculate the extrapolation factor for all road traffic accident victims admitted to hospital.

Then the extrapolation factors were calculated for two types of patient for whom the registration rate is extremely dependant: viz. the victim's modal split and the province in which the accident occurred.

The results, based on the databases for 1992 and 1993, are as follows: the number of hospitalised accident victims which the police recorded (an average of 12,065 per year) has to be multiplied by a factor of 1·64 in order to approximate the actual number (an average of 19,745 per year). This is approx. 2% more than the usual NPR number.
The extrapolation factors for the modal split range from 1·3 for car occupants to 2·5 for cyclists. The extrapolation factors for the province vary from 1·3 for Zeeland to 2·1 for Friesland. These factors apply for the years researched, but it may be assumed that these do not differ much from year to year. The factors found may therefore be used for a number of years. It is however necessary to repeat the linkage with a certain frequency, to be able to calculate changes.

The knowledge obtained about the coding mistakes found in this project has been collected together in a catalogue. Recommendations have been made for improving the quality of both databases.

The project has also lead to the conclusion that, under certain circumstances, the linkage method developed can also be used for linking other databases. This applies especially to databases where a lot of variable per record are available, but not all variables are recorded completely and faultlessly.

Inhoud

Voorwoord

Het voor u liggende rapport heeft een lange geschiedenis. De SWOV heeft al meer dan tien jaar geleden gepubliceerd over de onderregistratie van verkeersongevallen en -slachtoffers. Nadat ook de Raad voor de Verkeersveiligheid het probleem van de onderregistratie aan de orde had gesteld, heeft de Tweede Kamer het onderwerp ter hand genomen en is er een Projectenprogramma gestart om de registratie van verkeersongevallen te optimaliseren, door gebruik te maken van de politieregistratie en andere registraties.
Vanuit de Adviesdienst Verkeer en Vervoer van Rijkswaterstaat is naar aanleiding hiervan het voorstel gedaan het verkeersongevallenbestand te koppelen aan andere bestanden volgens een nieuwe methodiek, namelijk door rekening te houden met fouten in registraties en deze fouten in termen van statistische kansen te vertalen.
De SWOV kreeg de uitdagende opdracht deze methode te ontwikkelen en op zijn bruikbaarheid te beoordelen. Dit was door het onvoorspelbare verloop dat projecten waarin iets nieuws bedacht en gerealiseerd moet worden nu eenmaal hebben, niet altijd eenvoudig.

Het resultaat ligt er nu en op basis hiervan is het daadwerkelijk mogelijk en verantwoord om tot een betere (hogere) schatting van het aantal verkeersslachtoffers te komen.

Inleiding en leeswijzer

Aanleiding voor het onderzoek
In 1993 is in de Tweede Kamer gevraagd naar de werkelijke omvang van de verkeersonveiligheid in Nederland, in reactie op berichten dat slechts een beperkt deel van de ongevallen wordt geregistreerd. Er is in grote lijnen wel bekend hoe de registratiegraad van verkeersslachtoffers varieert over de verschillende categorieën verkeersdeelnemers, maar dit is mede gebaseerd op veronderstellingen en beperkte steekproeven. Inzicht in de exacte omvang van het aantal verkeersslachtoffers ontbreekt.
Onder de noemer 'Het topje van de ijsberg?' is het Projectenprogramma Registratie Verkeersongevallen (PPRV) gestart, dat de basis moet zijn voor verbetering van de registratie (Derriks & Driessen, 1994). De vraag naar de registratiegraad is gerelateerd aan de Strategische projecten in het onderzoeksprogramma waar, ten aanzien van andere bestanden dan het ongevallenbestand van de Hoofdafdeling Basisgegevens van de Adviesdienst Verkeer en Vervoer (AVV/BG), drie vragen centraal staan:
- welke aanvullende informatie bieden ze ten opzichte van de registratie door de politie?
- welke mogelijkheden bieden ze om extra inzicht te krijgen in de kwaliteit van de door de politie verstrekte informatie?
- welke informatie geven ze over de volledigheid van de registratie en welke ophoogfactoren kunnen worden bepaald op basis ervan?

Reikwijdte van het onderzoek en opbouw van de rapportage
De AVV heeft de SWOV opgedragen een onderzoek uit te voeren ter beantwoording van deze vragen. De nadruk ligt daarbij op de laatste vraag, die is toegespitst op de verkeersslachtoffers die in een ziekenhuis zijn opgenomen (de doelpopulatie) en die is beantwoord met behulp van de Landelijke Medische Registratie (LMR) van de Stichting Informatievoorziening Gezondheidszorg (SIG) als 'controlebestand'. Het onderzoek behelsde primair het schatten van de werkelijke omvang van het aantal in een ziekenhuis opgenomen slachtoffers, op basis van een koppeling van records uit de VOR en de SIG, waarbij de records niet op alle koppelkenmerken volledig identiek hoefden te zijn.

Dit rapport beschrijft de gevolgde procedure, de ontwikkelde methoden en de uitkomsten van het onderzoek; het bestaat uit drie delen, die corresponderen met de subfasen A, B en C waarin het onderzoek is ingedeeld.
Deel A beschrijft de wijze waarop de koppeling van de bestanden heeft plaatsgevonden. In deel B wordt, op basis van de gekoppelde bestanden, een schatting gemaakt van de werkelijke omvang van het aantal in een ziekenhuis opgenomen slachtoffers. Om dit te kunnen doen, is een nieuwe methode ontwikkeld, de 'footprint-methode', die apart behandeld wordt in § 3.4. Ten slotte worden in hoofdstuk 4 de ophoogfactoren berekend.
In deel C staan de algemene conclusies en aanbevelingen.

Deel A: De koppeling
Deel A behelst het koppelen van gegevens uit de bestanden van enerzijds de AVV/BG en anderzijds de SIG, alsmede het vaststellen van de mate waarin dit koppelen succesvol verlopen is. Bij het koppelen is een bepaalde foutenmarge toegelaten, een 'afstand' tussen records, omdat in iedere registratie

fouten kunnen voorkomen en omdat, als met deze fouten geen rekening wordt gehouden, records ten onrechte niet aan elkaar worden gekoppeld.

Deel A begint met een nadere uitwerking van de koppeltheorie (hoofdstuk 2). Tevens worden de in het onderzoek betrokken bestanden beschreven evenals de te kiezen koppelvariabelen (hoofdstuk 3 t/m 5).
Omdat tijdens het onderzoek bleek dat gedetailleerde informatie over de foutenbronnen en hun omvang onvoldoende aanwezig was (hoofdstuk 6), is een deelonderzoek ingelast, waarin door handmatige vergelijking en koppeling van records uit beide bestanden zicht ontstond op de fouten in beide bestanden (hoofdstuk 7). Op basis van deze informatie is vervolgens een afstandsfunctie bepaald (hoofdstuk 8) en een procedure opgesteld om de feitelijke koppeling geautomatiseerd te laten uitvoeren door een computer (hoofdstuk 9).
De koppeling is uitgevoerd en de uitkomsten zijn beoordeeld op hun bruikbaarheid (hoofdstuk 10). De gekoppelde records zijn ingedeeld naar koppelkwaliteit op grond van hun afstand tot het record waarmee ze zijn gekoppeld en de selectiviteit van de koppeling. Een koppeling is kwalitatief goed indien die afstand klein is en selectief als de afstanden tot alle andere records uit het andere bestand duidelijk groter zijn. Naarmate de koppelkwaliteit beter is, is aannemelijker dat gekoppelde paren uit beide bestanden daadwerkelijk hetzelfde slachtoffer betreffen.
De koppeling kan als geslaagd worden beschouwd (hoofdstuk 11). Dit is van belang, omdat naarmate de koppeling succesvoller verloopt, meer en betere informatie wordt gegenereerd over de kwaliteit van de bestanden VOR en LMR, met name over het optreden van fouten, omissies en 'vervuiling' door records die niet in de bestanden thuishoren.

Deel B: Schatting werkelijke omvang en bepaling ophoogfactoren
De ophoogfactoren zelf komen in deel B aan de orde. Ze zijn berekend op basis van een uitgebreide analyse van het resultaat van de koppeling in deel A, met behulp van een voor dit doel ontwikkelde methode. Bij het lezen van deel B wordt kennis van deel A verondersteld.
Ten behoeve van een goed begrip van deel B staan de definities van de in het onderzoek betrokken categorieën slachtoffers en de terminologie waarin de kwaliteit van de koppeling wordt beschreven, in een afzonderlijke paragraaf (§ 1.1).
Hoofdstuk 2 beschrijft de kwaliteitscontrole. De voorlopige analyse (§ 2.2) diende er in de eerste plaats toe na te gaan of er indicaties waren voor problemen in de koppelresultaten die het bepalen van ophoogfactoren in de weg zouden staan. Vervolgens is de kwaliteit nader beoordeeld op basis van een groot aantal tabellen, waarvan beknopt verslag wordt gedaan in § 2.4. Allereerst zijn de uitkomsten voor 1992 en 1993 vergeleken. De grote mate van overeenkomst versterkt het vertrouwen in de bruikbaarheid van de methode en de betrouwbaarheid van de uitkomsten. Op grond hiervan is ook besloten de twee analysejaren samen te voegen.Dat betekent dat vanaf hoofdstuk 3 alle gepresenteerde tabellen gegevens bevatten van de jaren 1992 en 1993 tezamen.
Vervolgens heeft een gedetailleerde analyse plaatsgevonden van de uitkomsten voor groepen uit het gekoppelde bestand die verschilden in koppelkwaliteit. Hierbij is onder andere gekeken naar de overeenkomst in de wijze van verkeersdeelname zoals gecodeerd in de VOR en de LMR. Deze variabele geeft een onafhankelijke controle op de koppeling, omdat de

vervoerwijze van het slachtoffer niet als koppelvariabele is gebruikt. Deze controle bevestigde de koppelresultaten.
Verder zijn het gekoppelde bestand en de restbestanden met elkaar vergeleken (§ 2.5). Hieruit bleek, zoals te verwachten was, dat het gekoppelde bestand relatief veel motorvoertuigongevallen bevatte en het LMR-restbestand juist relatief veel ongevallen waarbij geen motorvoertuig was betrokken.
Een eerste schatting van het aantal slachtofferrecords die in beide bestanden voorkomen, is gemaakt door het bepalen van de aantallen terecht gekoppelde records in de onderscheiden koppelkwaliteitsklassen (hoofdstuk 3). De hierbij gevolgde werkwijze wordt samengevat in § 3.3. Onderdeel hiervan is de ontwikkeling van een nieuwe methode, de footprint-methode, die in § 3.4 apart wordt beschreven.
Allereerst is aannemelijk gemaakt dat in de klasse met de hoogste koppelkwaliteit het aandeel onterecht gekoppelde records zeer klein is § 3.5). Vervolgens is voor deze koppelkwaliteitsklasse het patroon van combinaties van vervoerwijzen volgens LMR en VOR vastgesteld: de footprint van terecht gekoppelde records. Deze footprint is gebruikt om vast te stellen welk deel van de gekoppelde records in de verschillende andere kwaliteitsklassen terecht gekoppeld zijn. Dit is gedaan voor ongevallen met motorvoertuigen (§ 3.5.1), ongevallen met overige voertuigen (§ 3.5.2), spoorwegongevallen, ongevallen buiten de openbare weg en een restgroep (§ 3.5.3). De definitieve schatting van het aantal gemeenschappelijke records (de doorsnede) staat in hoofdstuk 4.
In aansluiting hierop zijn ook de aandelen doelpopulatie in de restbestanden geschat, evenals het aantal slachtoffers dat in geen van beide bestanden voorkomt (§ 4.3 t/m § 4.5). Hieruit volgt de factor waarmee het aantal slachtoffers waarvan de politie opgeeft dat ze in een ziekenhuis zijn opgenomen, moet worden vermenigvuldigd om de totale omvang van de doelpopulatie te krijgen. Bovendien zijn ophoogfactoren berekend naar wijze van verkeersdeelname (§ 5.2) en provincie (§ 5.3), omdat dit voor het beleid relevante variabelen zijn en de ophoogfactoren sterk verschillen naar vervoerwijze en provincie.
Deel B wordt afgesloten met aanbevelingen voor het beleid.

Deel C: Foutencatalogus en algemene conclusies
In deel C worden allereerst de in beide bestanden aangetroffen typen fouten met hun omvang gepresenteerd, zoals deze tijdens het hele onderzoekstraject zijn gebleken (hoofdstuk 2). Vervolgens zijn de voorwaarden behandeld waaronder de hier ontwikkelde koppelmethode op andere bestanden toegepast kan worden (hoofdstuk 3). De - mede door de koppeling ontstane - meerwaarde van het LMR-bestand ten opzichte van het VOR-bestand wordt toegelicht (hoofdstuk 4), gevolgd door aanbevelingen voor de verbetering van de kwaliteit van beide registraties (hoofdstuk 5). Besloten wordt met de aanbeveling de koppeling regelmatig te herhalen om veranderingen in de registratiegraad te kunnen volgen.

Deel A: De koppeling

1. Inleiding

In principe gaat het bij dit onderzoek om gewonde personen die door een verkeersongeval in een ziekenhuis zijn opgenomen. Deze groep gewonden valt onder de definitiegroep van beide hier beschouwde registraties: het verkeersongevallenbestand van AVV/BG, hierna te noemen VOR-bestand of VOR, en het ziekenhuisopnamenbestand van de SIG, hierna te noemen LMR-bestand of LMR. Voor beide geldt echter dat ze een veel grotere groep omvatten.

De twee registraties LMR en VOR geven elk veel, maar eenzijdige informatie. Door ze te koppelen kan op twee terreinen winst geboekt worden.

Ten eerste kan een veel betere indruk verkregen worden van de totale omvang van het aantal ernstig gewonden, zodat ook bekend is hoeveel de aantallen relevante gebeurtenissen uit elk der registraties opgehoogd moet worden om het totaal te verkrijgen. In het ideale geval leidt het onderzoek tot een ophoogfactor per combinatie van kenmerken van het slachtoffer, het ongeval en/of de locatie, zodat voor elke relevante doorsnijding van het ongevallenbestand een ophoogfactor bekend is. Het is wellicht te overwegen om de resultaten te vertalen naar een ophoogfactor per record; op die manier kunnen opgehoogde tabellen gemakkelijker uit het bestand worden verkregen. Dit zouden onderwerpen voor vervolgonderzoek kunnen zijn.

Ten tweede wordt zo duidelijk hoe het LMR- en het VOR-bestand elkaar kunnen aanvullen en wat de bruikbaarheid van het gekoppelde bestand is. Dit bestand is - bij gebleken geschiktheid - ideaal voor onderzoek naar de relaties tussen weg- en verkeerskenmerken van ongevallen enerzijds en medische gevolgen bij slachtoffers anderzijds. Als men ook de relaties met voertuigkenmerken wil onderzoeken is koppeling met bestanden als die van de Rijksdienst voor het Wegverkeer nodig.

1.1. De registratiegraad

Op basis van de resultaten van een koppeling kunnen ophoogfactoren worden berekend; een ophoogfactor is immers de inverse van een registratiegraad. Op zijn minst is een ophoogfactor voor alle in Nederlandse ziekenhuizen opgenomen verkeersslachtoffers gewenst als eindresultaat, met een nauwkeurigheid op basis waarvan beleidsuitspraken mogelijk zijn. Voor het bepalen van de totale omvang is een nauwkeurigheid van minstens 5% gewenst.
Voor het beoordelen van jaarlijkse verschillen moet de nauwkeurigheid het liefst niet te veel afwijken van de grootte-orde van de statistische fluctuaties, die ruim 1% bedragen bij de aantallen in ziekenhuizen opgenomen gewonden. Dit laatste zal zeer moeilijk bereikbaar zijn.
Voorts is gezocht naar significante afwijkingen van de ophoogfactoren, indien de slachtoffers worden onderverdeeld naar een aantal relevante kenmerken. De verschillen in ophoogfactoren naar onderverdelingen geven tevens inzicht in de variabelen die de afwijkende registratiegraad bepalen.

Bij de onderverdelingen is een minder grote nauwkeurigheid van de ophoogfactoren onvermijdelijk, mede doordat dan gewerkt wordt met kleine(re) aantallen.

1.2. Het gekoppelde bestand

AVV heeft tijdens voorbereidende besprekingen de voorkeur geuit voor methode van onderzoek die zoveel mogelijk uniforme is en die op zijn minst in theorie algemeen toepasbaar is op het ongevallenbestand van BG enerzijds en andere bestanden anderzijds. Het gaat hierbij om het koppelen op slachtoffer- of ongevalsniveau.
Deze methode heeft behalve de uniformiteit, als voordelen dat elke gewenste doorsnijding van het gekoppelde bestand is te maken en dat er wellicht ook een schatting kan worden gemaakt van het aantal slachtoffers dat in geen van beide bestanden voorkomt.

De resultaten van de koppeling leiden minstens tot:
- een oordeel over de geschiktheid van de methode van koppelen op slachtofferniveau, gebaseerd op een uitgevoerde koppeling, vooral wat de LMR betreft, maar ook een indicatie wat betreft de overdraagbaarheid van de methode naar andere bestanden zoals die van de verzekeringsmaatschappijen;
- uitspraken over de registratiegraad van in ziekenhuizen opgenomen verkeersslachtoffers;
- aanbevelingen ten aanzien van de bruikbaarheid van de LMR in termen van aanvullende of vervangende informatie en eventuele wenselijke verbeteringen van de LMR;
- aanbevelingen om de VOR-registratie te verbeteren.

Ook wordt zo duidelijk welke kennis kan worden verkregen:
- uit de afzonderlijke bestanden;
- door vergelijking van de bestanden;
- over wat toegevoegd wordt door een eenmalige koppeling;
- over wat toegevoegd wordt door een herhaling.

Zo kan beoordeeld worden of in de toekomst vaker gekoppeld moet worden, en met welke tussenpozen. Misschien is het zelfs aan te bevelen de koppeling jaarlijks uit te voeren.

1.3. Opzet

In dit deel worden de gegevens gekoppeld en wordt de mate waarin dit slaagt vastgesteld. Een nieuwe wijze van koppelen is ontwikkeld en toegepast op een ruim gekozen deel van het LMR-bestand en het volledige VOR-bestand. Het doel is om zo veel mogelijk records aan elkaar te koppelen waarvan het aannemelijk is dat ze hetzelfde slachtoffer betreffen, zonder het aantal ten onrechte gekoppelde records te groot te maken.

1.4. Resultaat

Het resultaat van dit deel van het onderzoek bestaat uit een voorlopige foutencatalogus op basis van interviews met de bestandsbeheerders, interviews over de aanwezige kennis bij de SWOV, een koppelprotocol, een

afstandsfunctie en de gekoppelde bestanden. Het koppelprotocol is ruimer toepasbaar, ten behoeve van andere te koppelen bestanden. De afstandsfunctie is een weerslag van de fouten zoals die in de registraties voorkomen. Daaruit zijn aanbevelingen te distilleren voor de verbetering van de bestanden.
De gekoppelde bestanden worden in deel B gebruikt om ophoogfactoren te berekenen, om in het vervolg uit de (door de bestandsbeheerders gerapporteerde) aantallen slachtoffers de werkelijke waarden te kunnen schatten.

2. De theorie van het koppelen

2.1. Inleiding

Meestal wordt onder het *koppelen* van bestanden verstaan dat records aan elkaar worden toegewezen op basis van een uniek kenmerk. Zo kan in Denemarken gekoppeld worden op basis van het Persoonsnummer, dat zowel door de politie als in het ziekenhuis geregistreerd wordt (Larsen, 1992).
In Nederland wordt zo'n kenmerk niet geregistreerd; er wordt wel gewerkt met een *koppelsleutel*, een combinatie van kenmerken die (nagenoeg) uniek zijn. Deze laatste koppelwijze wordt ook wel *matching* genoemd. Wij zullen het toch in het vervolg over koppelen hebben, behalve als misverstand mogelijk is.
In 1987 is een proefkoppeling over het jaar 1985 uitgevoerd door de SIG, in samenwerking met de SWOV. De resultaten zijn gerapporteerd door de SIG (Nauta, 1988) en de SWOV (Blokpoel & Polak, 1991). Uit de resultaten van de proefkoppeling over 1985 is gebleken dat de toen gebruikte combinatie van vier kenmerken (geboortedatum, ziekenhuis, datum ongeval/opname en geslacht) inderdaad praktisch uniek is. Dit betekende dat (met grote waarschijnlijkheid) heel weinig records ten onrechte gekoppeld zijn, maar dat veel records ten onrechte *niet* gekoppeld werden, door onnauwkeurigheden die exacte overeenstemming verstoorden. Om die reden is het zinvol te onderzoeken hoe - gegeven het uiteindelijke doel van de koppeling - de 'beste' koppelwijze en -sleutel vastgesteld dient te worden, indien 'kleine' onnauwkeurigheden in de gegevens worden toegelaten. Daartoe moet eerst een theoretisch kader geformuleerd worden, waarbij de afwijkingen en de fouten in de te koppelen bestanden een essentiële rol spelen. Immers, bij foutloze bestanden is een perfecte koppeling mogelijk.

2.2. Doel

De VOR-LMR-koppeling dient twee onderscheiden doelen:
- Het bepalen van de omvang van het totale aantal ernstig gewonden, en het bepalen van onderverdelingen, zodat ophoogfactoren kunnen worden vastgesteld.
- Het mogelijk maken van onderzoek naar de relaties tussen ongevals- en letselkenmerken.

Het ligt voor de hand dat beide doelen niet dezelfde eisen aan de koppeling hoeven te stellen. Waar het in het vervolg gaat om het koppelen van records met minder dan de hoogste graad van waarschijnlijkheid zal dat vooral van belang zijn voor het eerste doel; voor het tweede doel is een (kleiner) gekoppeld bestand van de hoogste kwaliteit optimaal, voor het eerste doel zal de grens verschoven worden in de richting van een zo groot mogelijk bestand van acceptabele kwaliteit.

2.3. De metriek van de ruimte opgespannen door de koppelkenmerken

Essentieel bij het koppelen is *de mate van overeenstemming van de (voor het aan elkaar toewijzen van twee records gebruikte) kenmerken.* Deze mate

van overeenstemming kan gevisualiseerd worden als een gegeneraliseerde afstand in een meer-dimensionale ruimte, de *koppelruimte*. De dimensies van die ruimte zijn de koppelkenmerken. Ieder record wordt gerepresenteerd door een *punt* in de koppelruimte. Alle - voorkomende of mogelijke - records vormen een discrete verzameling punten in die ruimte, door de digitale verwerking in het record. Twee punten kunnen samenvallen, maar als ze niet samenvallen hebben ze een zekere, eindige afstand. De wijze van berekenen van die afstand wordt de *metriek* van die ruimte genoemd.

De afstand in de koppelruimte tussen twee punten uit elk der bestanden moet een directe maat zijn van de *waarschijnlijkheid* dat die twee punten uit dezelfde gebeurtenis voortkomen, in de zin dat een grotere afstand een kleinere waarschijnlijkheid betekent.

Bij een koppeling tussen twee complete en foutloze registraties zullen de punten altijd in paren voorkomen, één uit elk der bestanden. We wijzen dan die paren, die afstand 0 hebben, aan elkaar toe. Wèl kunnen in de afzonderlijke bestanden *(administratieve) meerlingen* voorkomen: dit zijn twee of meer records *in hetzelfde bestand* waarvan alle koppelkenmerken dezelfde waarde hebben. Bij foutloze bestanden zullen die in beide bestanden zitten en niet ondubbelzinnig gekoppeld kunnen worden zonder extra informatie (als een uniek persoonsnummer) te gebruiken.
Zo'n administratieve tweeling kan een 'echte tweeling' zijn: een tweeling van gelijk geslacht heeft een gezamenlijk ongeval en wordt ook samen opgenomen. Maar het kunnen ook twee mannen zijn die toevallig op dezelfde dag geboren zijn en op dezelfde dag een ongeval hadden waarbij ze in hetzelfde ziekenhuis werden opgenomen.
Het is mogelijk om het optreden van deze meerlingen tot een minimum te beperken door de koppelsleutel selectief te maken. In termen van de koppelruimte: we maken het aantal mogelijke punten in de koppelruimte zo groot dat de dichtheid in die ruimte - het aantal records gedeeld door het aantal punten - zo laag is dat meerlingen (praktisch) niet meer voorkomen. Omdat de resolutie van de kenmerken gegeven is, kan het aantal mogelijke punten alleen vergroot worden door kenmerken toe te voegen aan de koppelsleutel. We kunnen het resultaat ook weergeven in termen van *afstand*. De afstand 0 moet tussen twee records binnen een te koppelen bestand (praktisch) niet voorkomen; bij het koppelen van twee records uit verschillende bestanden wijzen we alleen paren met afstand 0 aan elkaar toe. Dit is de simpelste metriek: we onderscheiden twee afstanden, 0 en (veel) groter dan 0, in termen van kansen: òf praktische zekerheid dat ze bij elkaar horen òf dat dit juist niet zo is.

2.4. De rol van fouten

In termen van de koppelruimte betekent een fout dat een punt op een verkeerde plaats zit, op enige afstand van zijn eigenlijke plaats. Als alleen punten met afstand 0 gekoppeld worden leidt deze fout tot het ten onrechte niet koppelen van twee bij elkaar horende records (en het mogelijk ten onrechte koppelen van de foutief geplaatste punt met een ander punt). Naarmate per (koppel)kenmerk meer fouten voorkomen en de koppelsleutel meer kenmerken omvat, zal het aantal fout geplaatste punten toenemen, waardoor het aantal toewijzingen vermindert. Als de aard van de fouten zodanig is dat een punt maar weinig verplaatst wordt, en de dichtheid van

de punten zo laag is dat door fouten verplaatste punten niet in de buurt van andere punten komen, kan toch een goede toewijzing volgen door bij het toewijzen een bepaalde afstand groter dan 0 toe te staan. In dat geval bestaat een door de foutenverdeling gegenereerde metriek in de koppelruimte die vertaald kan worden in een op afstand gebaseerd toewijzingsprotocol.
Om de fouten en de bijbehorende metriek te weten te komen, is kennis nodig over de foutenverdelingen van de koppelkenmerken. Deze kennis is nog in onvoldoende mate voorhanden; dit vormt dan ook een van de onderwerpen van dit onderzoek.

2.5. De rol van onbekenden

Kenmerken die als 'onbekend' zijn gecodeerd vormen een speciaal probleem. Dit kan echter met het afstandsbegrip aangepakt worden. Het is duidelijk, dat als twee punten samenvallen, maar met beide de waarde onbekend voor een der kenmerken, hun afstand formeel 0 is, wat niet betekent dat ze zonder meer aan elkaar toegewezen kunnen worden.

Bij de proefkoppeling van 1987 zijn de records met 'onbekend' als waarde van een der koppelkenmerken van te voren verwijderd. Dit gebeurde omdat de proef relatief eenvoudig gehouden moest worden. Bij de hier beschreven koppeling kunnen we deze records meenemen omdat extra informatie uit aanvullende kenmerken toch kan leiden tot een betrouwbare toewijzing. Daarbij speelt de kans dat de variabele met de waarde onbekend in werkelijkheid dezelfde waarde heeft als die in het andere bestand een rol. Onderscheiden moet worden het geval waarbij de variabele in één der bestanden onbekend is en die waarbij dezelfde variabele in beide bestanden onbekend is. In het eerste geval wordt een bepaalde afstand tussen 'onbekend' en elk der bekende waarden van het kenmerk ingevoerd. Dezelfde afstand kan gehanteerd worden als in beide bestanden hetzelfde kenmerk onbekend is, omdat de kans dat het om dezelfde waarde gaat gelijk is in de twee hierboven genoemde gevallen.
We kunnen ook voor verschillende waarden van de bekende waarde verschillende afstanden toekennen tot de onbekende, bijvoorbeeld om in rekening te brengen dat het aannemelijker is dat de onbekende waarde een veel voorkomende is. Daarvoor is echter kennis over die verdeling nodig.

2.6. Het meetniveau van het kenmerk

Bij het toekennen van de afstanden spelen twee aspecten een rol: het meetniveau van het kenmerk en de aard van het proces dat de fouten in de kenmerken veroorzaakt. Bij een *nominaal* meetniveau kan een fout in principe tot elke andere waarde van het kenmerk leiden, weliswaar met mogelijk verschillende kansen. Bij een *ordinaal* meetniveau kan het voorkomen dat de kans kleiner wordt naarmate de fout een verschuiving van meer stappen langs de schaal veroorzaakt. Bij *interval-* en *ratioschalen* zal het vaker voorkomen dat van een metrische afstand gesproken kan worden. Essentieel is dus de door de fouten bij het meetproces van de koppelkenmerken gegenereerde metriek in de koppelruimte. Deze metriek kan ook opgevat worden als een weging: kenmerken die met relatief veel fouten behept zijn, dragen weinig bij tot de afstand; praktisch foutloze kenmerken dragen juist veel bij tot de afstand, zodat de laatste een groter gewicht hebben.

3. De bestanden

Het LMR-bestand omvat (nagenoeg) alle personen die in Nederlandse ziekenhuizen zijn opgenomen. Het bestand is - jaarlijks - georganiseerd op basis van *ontslagen* patiënten. Als iemand als gevolg van hetzelfde ongeval verscheidene keren wordt opgenomen (en dus ook ontslagen), komt hij of zij er dus evenveel malen in voor. Het LMR-bestand is zeer groot, de opnamen met als kenmerk *vervoersongeval* bedragen jaarlijks ongeveer 19.000.

Het VOR-bestand bevat (nagenoeg) alle verkeersslachtoffers waarvan de politie kennis heeft genomen en een formulier heeft gestuurd aan AVV/BG. Het bestand is hiërarchisch opgebouwd, met als structuur de *locatie,* waarop *ongevallen* gebeuren, waarbij *objecten* betrokken zijn, waar *slachtoffers* bij kunnen horen. In de periode 1992-1993 kende het VOR-bestand jaarlijks circa 49.000 verkeersslachtoffers waarvan ongeveer 12.000 (volgens opgave van de politie) in een ziekenhuis waren opgenomen.

3.1. De onderzoeksperiode

De ervaring met een eerdere proefkoppeling over het jaar 1985 (Blokpoel & Polak, 1991) heeft geleerd dat gegevens over één kalenderjaar al geschikt zijn voor onderzoek naar bepaalde aspecten. Omdat het hier vooral gaat om de registratiegraad van onderverdelingen van de VOR-gegevens en het bekend is dat de totale registratiegraad jaarlijks wisselt en tevens een dalende tendens vertoont, is in overleg met SIG en AVV besloten meer jaren, met name 1992 en 1993, te onderzoeken.
Omdat ook bleek dat - door de zeer sterk toegenomen computercapaciteit sinds de vorige koppeling - de omvang van de te verwerken bestanden geen probleem meer vormde, kan ook het jaar 1994 worden toegevoegd. Het was echter bij de opzet van dit onderzoek onzeker of dit laatste jaar ook in feite onderzocht zou kunnen worden, omdat dit afhangt van het beschikbaar komen van SIG-gegevens over (het eerste deel van) 1995; immers het SIG-bestand is een bestand van ontslagen patiënten en in 1994 opgenomen personen zullen voor een deel pas in 1995 (of nog later) uit het ziekenhuis ontslagen worden. De verwachting was dat deze gegevens in het voorjaar van 1996 beschikbaar zouden zijn, maar dit bleek niet het geval te zijn.
De onderzoeksperiode bestaat dus uit de jaren 1992 en 1993. Omdat de VOR-gegevens over 1994 bij de SWOV klaar liggen is het desgewenst eenvoudig dit jaar in een later stadium toch te analyseren.

3.2. De selectie uit het LMR-bestand

Uit het LMR-bestand moet de - relatief kleine - deelverzameling van verkeersslachtoffers geselecteerd worden. In principe kan dit met de E-code. Dit is een aan de (internationale) Classificatie van Ziekten (Classificatie van Ziekten, 1980) toegevoegde codering die verplicht is als de opname in een ziekenhuis het gevolg is van een ongeval of vergiftiging. De E-code bestaat uit de letter E, gevolgd door drie cijfers, een punt en dan nog één of twee cijfers. De cijfers vóór de punt geven het soort ongeval aan.

De internationale definitie van een verkeersongeval luidt:

Een gebeurtenis op de openbare weg, die verband houdt met het verkeer, waarbij minstens één rijdend voertuig is betrokken en ten gevolge waarvan één of meer weggebruikers zijn overleden en/of gewond.

Het CBS en AVV/BG stellen vanzelfsprekend ook de eis dat het ongeval in Nederland gebeurd moet zijn.

De systematiek van de E-code kent de groep *vervoersongevallen*, waaronder behalve verkeersongevallen ook ongevallen met vlieg- en vaartuigen vallen. De verkeersongevallen die voldoen aan de internationale definitie vormen helaas geen aparte deelgroep in de systematiek. Omdat het niet mogelijk is (zie ook verder) de verkeersongevallen ondubbelzinnig te selecteren, is er de voorkeur aan gegeven de selectie ruim te houden, in die zin dat codes die (naast andere) waarschijnlijk ook (enkele) verkeersongevallen bevatten ook in de selectie meegenomen worden. Alleen op die manier kan kennis verkregen worden over de aantallen waar het om gaat. Onderscheiden worden, in volgorde van overeenstemming met de definitie:

E810-E819: *Verkeersongevallen met een motorvoertuig* (op de openbare weg).
Hieronder vallen ook botsingen (van motorvoertuigen) met een trein; brom- en snorfietsen worden ook tot de motorvoertuigen gerekend.

E826-E829: *Ongevallen met andere wegvoertuigen.*
Hier wordt niet de eis gesteld dat het ongeval op de openbare weg gebeurd moet zijn.

Deze twee groepen worden in het vervolg tezamen de standaardgroep genoemd, omdat ze de gebruikelijke basis vormen voor de presentatie van LMR-cijfers over in ziekenhuizen opgenomen verkeersslachtoffers.

Omdat de binnen de LMR gebruikte definitie van 'openbare weg' bepaalde gebieden niet omvat die volgens de codeerinstructie van AVV (AVV/BG, 1993) wèl meegenomen worden, zoals vliegveldterreinen en parkeerplaatsen, zullen de ongevallen met motorvoertuigen buiten de openbare weg ook bij de selectie gevoegd worden:

E820-E825: *Niet-verkeersongevallen met een motorvoertuig.*
Hieronder zullen ongevallen voorkomen die inderdaad geen verkeersongevallen zijn.

De systematiek van de E-code heeft tot gevolg dat onder de (zeldzame) ongevallen met overlevenden, waarbij een trein in botsing komt met een ander wegvoertuig (anders dan de bij E810-E819 bedoelde motorvoertuigen) of een voetganger, verkeersongevallen kunnen voorkomen. Daarom voegen we nog toe:

E801: *Spoorwegongeval door botsing met een ander object.*
E805-E807: *Geraakt door rollend materieel, Overige gespecificeerde spoorwegongevallen en Spoorwegongeval van niet-gespecificeerde aard.*
Ook hier wordt de eis van gebeuren op de openbare weg niet gesteld, zodat te veel wordt meegenomen.

De E-code kent ook *Niet gespecificeerde ongevallen*:

E928.9; hierbij wordt door een tweede cijfer achter de punt de plaats van het ongeval aangegeven: de cijferwaarden 0, 4, 5, 6, 8 en 9 zijn geselecteerd.
.0: *In en rondom huis*. Hieronder valt ook *Erf* en *Oprit*.
.4: *Plaats voor recreatie of sport*. Onder meer *Openbaar park*.
.5: *Straat of andere openbare weg*.
.6: *Openbaar gebouw*. Ook *Markt* en *Vliegveld* valt hieronder.
.8: *Andere gespecificeerde plaatsen*. Zoals *Openbare plaats NNO* en *Parkeerplaats en -terrein*,
.9: *Niet gespecificeerde plaats*. 'Onbekend'.

Daarmee zijn de openbare weg en andere bij de VOR meegenomen locaties in ieder geval meegenomen. Het gaat hier om een groot aantal records: in 1992 waren het er bijna 7.000.

Een zelfmoord(poging) in het verkeer hoort volgens de codeerinstructie van AVV/BG niet tot de verkeersongevallen; als de politie vaststelt dat het om zelfmoord gaat, komt het dus niet in de VOR-registratie terecht.
Ook bij de LMR valt zelfmoord onder een andere code. Omdat het vaak moeilijk of niet vast te stellen is, zowel voor de politie als voor verplegend personeel, of inderdaad sprake is van (een poging tot) zelfmoord, en - om redenen van privacy-bescherming - ook van kennis op dit gebied geen melding gedaan zou kunnen worden, is besloten de code voor zelfmoord-(poging) toe te voegen. De wijze waarop de poging is ondernomen is gecodeerd in het cijfer achter de punt en daarvan komen in aanmerking:

E958: *Zelfmoord en zelf toegebracht letsel door andere en niet gespecificeerde middelen,*
.0 *Voor een bewegend voorwerp springen of liggen;*
.5 *Te pletter rijden met een motorvoertuig;*
.8 *Overige gespecificeerde middelen;*
.9 *Niet gespecificeerd middel.*

Ten slotte is ook toegevoegd:

E988: *Letsel door andere en niet gespecificeerde middelen, waarvan niet vastgesteld is of dit opzettelijk of niet opzettelijk is toegebracht.* Hierbij worden dezelfde cijfers achter de punt toegepast als bij E958 (zelf-moord).

In totaal bestaat het te koppelen jaarbestand uit ongeveer 26.000 records.

3.3. De selectie uit het VOR-bestand

Hoewel in dit bestand een code voorkomt die aangeeft of en in welk ziekenhuis een slachtoffer is opgenomen is het toch nodig geoordeeld om alle verkeersslachtoffers in het te koppelen bestand op te nemen.
Ten eerste is bij de proefkoppeling van 1987 gebleken dat onder de slachtoffers die volgens opgave van de politie wel vervoerd waren naar een ziekenhuis maar aldaar niet opgenomen, circa 10% toch te matchen waren met het LMR-bestand.
Ten tweede geeft de codeerinstructie aan dat als bekend is dat een slachtoffer *later* is opgenomen dit gecodeerd moet worden als *Niet opgenomen.* Hier wordt onder *later* verstaan dat het slachtoffer niet direct van de plaats van het ongeval is vervoerd naar een ziekenhuis.
Ten derde is het aannemelijk dat de politie in veel gevallen waarbij een slachtoffer later in een ziekenhuis wordt opgenomen daarvan niet op de hoogte kan zijn, zodat de registratie wel onjuist moet zijn. Geconcludeerd kan worden dat de code *Opgenomen in een ziekenhuis* in de meeste gevallen alleen in het record wordt opgenomen als dat door eigen waarneming van de politie is vastgesteld èn het vervoer naar het ziekenhuis direct volgde op het ongeval.
Dit alles heeft ertoe geleid dat behalve de opgenomen gewonden ook de 'overige' gewonden in het koppelbestand zijn meegenomen. In 1992 ging het om 12.108 opgenomen gewonden, 3.327 slachtoffers waarvan onbekend was of ze opgenomen zijn en 33.926 slachtoffers die volgens de politie niet zijn opgenomen. Daaronder vallen 741 slachtoffers die ter plaatse van het ongeval zijn overleden, 16.727 slachtoffers die wel naar een ziekenhuis vervoerd zijn en 16.458 slachtoffers die niet naar een ziekenhuis vervoerd zijn.

3.4. Verwerving en voorbewerking

3.4.1. *Het LMR-bestand*

Uit de beschikbare, op jaar van *ontslag* uit het ziekenhuis geordende, bestanden zijn op jaar van *opname* gebaseerde bestanden aangemaakt. Daarbij is voor het opnamejaar 1992 gebruikgemaakt van de ontslagjaren 1992 t/m 1994 en voor opnamejaar 1993 de jaren 1993 en 1994. Omdat het uiterst zelden voorkomt dat verkeersslachtoffers twee kalenderjaren later ontslagen worden dan opgenomen, worden zo hoogstens enkele records van opnamejaar 1993 gemist.
Records zijn verwijderd die vervolgopnamen van het zelfde slachtoffer betreffen. De voor SWOV-gebruik ontwikkelde programmatuur, die de door sommige ziekenhuizen nog gebruikte oude E-codes omzet naar nieuwe E-codes, is daarna op de bestanden toegepast.
De verwerving en bewerking zijn een interne SIG-zaak, die aldaar geregeld is.

3.4.2. *Het VOR-bestand*

Het is gebleken dat de bij de SWOV aanwezige VOR-bestanden alle gegevens bevatten die voor de koppeling nodig zijn. De verwerving was dus een interne SWOV-zaak. Wel bleek dat voor 1992 een klein deel van de

gegevens (een dertigtal ontbrekende geboortedata) aangevuld diende te worden.

Omdat de codering van ziekenhuizen bij de VOR anders is dan de officiële codering die de SIG hanteert, namelijk de Instellingenlijst Gezondheidszorg (SIG, 1995), moesten drie transpositietabellen opgesteld worden, voor de jaren 1992, 1993 en 1994. De jaarlijkse veranderingen reflecteren het voortdurende proces van fusies tussen ziekenhuizen. Met behulp van deze tabellen is de code aan het VOR-bestand toegevoegd.

Uit de twee variabelen 'Datum ongeval' en 'Tijdstip ongeval' is de variabele Epoch samengesteld, die datum en tijdstip combineert.

4. De afwijkingen

In beginsel is het streven bij deze koppeling om die delen uit de beide bestanden te selecteren die slachtoffers van verkeersongevallen in Nederland betreffen die in een ziekenhuis zijn opgenomen (de *doelpopulatie*), zoveel mogelijk onder uitsluiting van slachtoffers die niet aan deze definitie voldoen. Daarbij is ervoor gekozen die deelpopulaties uit de bestanden toe te voegen waaronder nog opgenomen verkeersslachtoffers zouden kunnen voorkomen. Dit heeft tot gevolg gehad dat beide bestanden een aanmerkelijk grotere omvang gekregen hebben dan zou volgen uit de gebruikelijke selectie. Omdat geen enkele registratie foutloos is zullen *afwijkingen* optreden: gevallen die ten onrechte in het bestand voorkomen ('te veel'), en gevallen die ten onrechte niet voorkomen ('te weinig').

4.1. **Het LMR-bestand**

De hierna genoemde percentages slaan terug op de - uit de gebruikelijke selectie van de standaardgroep (E-code 810-819 en 826-829) volgende - omvang van het LMR-bestand van circa 19.000. (Hier zijn de eerder genoemde additionele E-codes zoals 'onbekend' niet bij inbegrepen.)

4.1.1. *Te weinig*

Er zijn bij besprekingen met SIG- en AVV/BG-functionarissen drie processen naar voren gekomen waardoor tot de doelpopulatie behorende slachtoffers niet in het LMR-bestand terecht komen.
Het eerste betreft slachtoffers van verkeersongevallen (in Nederland) die in een buitenlands ziekenhuis worden opgenomen. Zo worden misschien wel circa 10% van de slachtoffers met brandwonden vanuit de regio Heerlen in een (gespecialiseerd) ziekenhuis in Aken opgenomen. Dit proces zal zich voornamelijk in grensstreken voordoen. In een later stadium zou onderzocht kunnen worden in welke mate zich dit voordoet, doordat de locatie van het ongeval bij de VOR bekend is.
Ook kunnen gevallen gemist worden door onderregistratie. Het onderregistreren kan ontstaan doordat opnamen niet aan de SIG gerapporteerd worden, bijvoorbeeld door te late inzending van gegevens, door het verloren gaan van gegevens of het bewust niet registreren van bepaalde patiënten (illegalen). De omvang van de onderregistratie is slecht bekend, maar zou bij enkele (grote) ziekenhuizen wel 1 à 2% kunnen bedragen.
Ten slotte zullen die slachtoffers gemist worden die tijdens het aanmaken van het bestand nog opgenomen zijn. Omdat opnameduren van meer dan een jaar zeer zeldzaam zijn en de analyse betrekking heeft op gegevens van meer dan een jaar geleden gaat het hier om één à twee gevallen per analysejaar.
Vóór 1992 bestond onderregistratie doordat niet alle ziekenhuizen waar ongevalspatiënten opgenomen kunnen worden aan de LMR meededen. Inmiddels doen volgens opgave van de SIG alle in aanmerking komende ziekenhuizen mee.

4.1.2. *Te veel*

Het is bekend dat het (in de grensstreek) niet zeldzaam is dat slachtoffers van verkeersongevallen in het buitenland in Nederlandse ziekenhuizen worden opgenomen. Dit is bijvoorbeeld het geval met slachtoffers met hersenletsel uit Duitsland die vaak in Heerlen worden opgenomen. Het kan ook gaan om Nederlanders die na een ongeluk in het buitenland liever naar Nederland vervoerd worden. Deze gevallen zijn niet in dit bestand te herkennen, maar zullen naar verwachting het meest voorkomen bij ziekenhuizen in de grensstreek. In deze ziekenhuizen zou het ook om vele procenten kunnen gaan. Doordat de locatie van alle ziekenhuizen bekend is kan dit probleem in principe onderzocht worden. Dit is in deze fase niet gebeurd.
Ten tweede zitten bij de geselecteerde gevallen ook slachtoffers van ongevallen die niet op de openbare weg gebeurd zijn. Deze zijn, zoals hierboven omschreven, meegenomen omdat ze gedeeltelijk niet herkenbaar zijn en bij de ongevallen met motorvoertuigen onder een andere definitie vallen dan bij AVV/BG. De omvang van deze groep bedraagt, naar wordt aangenomen op basis van eerder onderzoek, circa 400 (2%) per jaar.
Doordat ook de E-codes voor ongevallen met niet-gespecificeerde oorzaken zijn toegevoegd zal het bestand vele duizenden te veel hebben die geen verkeersongeval waren.
Er is gepoogd opnamen van slachtoffers die eerder voor hetzelfde ongeval waren opgenomen uit het bestand te halen. Zonder die schoning zou het om hoogstens 3% 'te veel' gaan, dit aantal is naar wordt aangenomen teruggebracht tot minder dan 0,5%.
Een laatste mogelijkheid is het optreden van meervoudige registratie van hetzelfde geval. Ondanks controle zouden dit soort administratieve meerlingen voor kunnen komen. Tijdens de handmatig gestuurde koppeling bleken inderdaad enkele dubbele registraties voor te komen. Deze zijn 'ontdubbeld'.

4.2. **Het VOR-bestand**

Hierna genoemde percentages hebben betrekking op alle verkeersslachtoffers in het VOR-bestand die als 'opgenomen in een ziekenhuis' zijn geregistreerd: circa 12.000.

4.2.1. *Te weinig*

De belangrijkste categorie is hier onderregistratie door niet aanwezig zijn van de politie, dan wel het niet insturen van een registratieformulier. Deze categorie bedraagt voor ziekenhuispatiënten, naar tot nu toe werd aangenomen, meer dan 30%! Een belangrijk doel van dit onderzoek is deze omvang beter te meten en in kaart te brengen.
Een klein deel van de formulieren die de politie wel heeft opgestuurd, komt niet in de registratie terecht doordat locatie-informatie ontbreekt of het formulier te laat wordt ingezonden ('naijlers'). Volgens opgave van AVV/BG gaat het hier om ruim honderd gevallen zonder locatie per jaar, en ruim 700 naijlers, van alle slachtoffers. Bij elkaar dus circa 2%.

4.2.2. *Te veel*

Doordat alle verkeersslachtoffers geselecteerd zijn, is het aantal 'te veel' aanzienlijk, naar schatting meer dan 30.000.
Ook hier bestaat de mogelijkheid van dubbele (of meervoudige) registratie van ongevalsslachtoffers. Bij de uitgebreidere controle van de dodelijke ongevallen moet geregeld 'ontdubbeld' worden. Het is dus niet onaannemelijk dat dubbelen bij de overige ongevallen (en dus ook bij de slachtoffers) door de controle zullen slippen. Uit een bij de SWOV uitgevoerde voorlopige analyse van het VOR-bestand van 1994 bleek het bestaan van zes paren slachtoffers met gelijke waarden voor geslacht, geboortedatum, datum en tijdstip van het ongeval en vervoer naar hetzelfde ziekenhuis. Deze 'administratieve meerlingen' kunnen reëel zijn, maar ook door dubbele registratie ontstaan zijn.

4.3. **Afwijkingen door de selectie**

Door de selectie van delen uit bestanden kunnen nieuwe afwijkingen ontstaan. Bij dit onderzoek zal het geselecteerde LMR-bestand die verkeersslachtoffers missen die geen E-code hebben gekregen uit de selectielijst. Volgens opgave van de SIG kan dit alleen door niet-opgemerkte codeerfouten gebeurd zijn en de kans daarop is door de kwaliteitscontrole van een zodanige orde dat niet meer dan circa 0,2% gevallen gemist worden. Op een totaal van 19.000 zou dit neerkomen op veertig gevallen.
Door de wijze van selecteren - maximaal ruim bij het VOR-bestand en ruim bij het LMR-bestand - zullen wel veel gevallen in de bestanden zitten die niet tot de doelpopulatie behoren.

5. De keuze van de koppelvariabelen

5.1. Inleiding

De variabelen die bij de koppeling een rol spelen, kunnen als volgt voor ieder der te koppelen bestanden onderscheiden worden:

- *Identificatievariabelen*: dit zijn een of eventueel meer variabelen die het slachtoffer binnen ieder bestand uniek aanduiden. Hierdoor kunnen later, na de koppeling, controles uitgevoerd worden en eventueel variabelen toegevoegd worden.

- *Koppelvariabelen*: Dit zijn de variabelen die betrokken worden bij het koppelprotocol. De te matchen records uit elk van de twee bestanden dienen zo min mogelijk te verschillen in de waarden van iedere koppel-variabele.

- *Controlevariabelen*: Dit zijn variabelen die gebruikt kunnen worden bij de (nadere) controle op de juistheid van de koppelingsprocedure.

- *Analysevariabelen*: Dit zijn de overige relevante variabelen.

Voor het uitvoeren van de koppeling zijn de identificatievariabelen en de koppelvariabelen absoluut nodig. De controlevariabelen zijn nodig voor de controle van de juistheid van de koppeling. Om de omvang van de bestanden te beperken is met de opname van deze variabelen in de records volstaan.

5.2. Identificatievariabelen

Behalve de na selectie en sorteerslag toegekende rangnummers zijn in beide bestanden ook de bestaande identificatievariabelen overgenomen. Bij de LMR gaat het om het Patiëntnummer en het Opnamenummer, bij de VOR om het VOR-nummer dat aan ieder *ongeval* wordt toegekend, en een bij de SWOV toegevoegd slachtoffernummer: *KEY_SLA*.

5.3. Koppelvariabelen

Dit zijn variabelen die òf gebruikt zijn voor de selectie, òf in principe voor alle records in elk bestand met goede kwaliteit zijn geregistreerd, en op (soort)gelijke wijze voorkomen:

- de E-code (LMR);
- de variabele ERNSTSL (VOR), die onder meer aangeeft of een slacht-offer is opgenomen;
- het geslacht;
- de geboortedatum;
- de Epoch: datum en het tijdstip van opname, respectievelijk ongeval;
- het ziekenhuis(nummer);

5.4. Controlevariabelen

Controlevariabelen zijn variabelen die uniek zijn voor elk der bestanden, variabelen die met mindere kwaliteit geregistreerd worden, dan wel variabelen die maar voor een deel van de records van toepassing zijn.

Voor het LMR-bestand gaat het om:
- de locatie van het ziekenhuis (postcode of gemeente);
- de woonlocatie van het slachtoffer (postcode of gemeente);
- de datum van overlijden (indien van toepassing);

Voor het VOR-bestand gaat het om:
- de vervoerwijze van het slachtoffer;
- de botspartner (=object of andere vervoermiddel waartegen gebotst is);
- bestuurder of passagier;
- slachtoffer opgenomen/niet opgenomen;
- slachtoffer vervoerd per ...;
- de gemeente van het ongeval;
- de maand van binnenkomst van het registratieformulier;
- de datum van overlijden (indien van toepassing).

6. De foutencatalogus, bestaande kennis

6.1. Inleiding

Deze foutencatalogus is gebaseerd op interviews van de beheerders en interviews over de aanwezige kennis bij de SWOV. Bij de opzet van dit onderzoek werd aangenomen dat aard en omvang van de fouten die in de loop van het registratieproces ontstaan, op deze wijze verkregen zouden kunnen worden. Het is echter gebleken dat minder kennis voorhanden is over de foutkansen van de koppelvariabelen dan gewenst en verwacht werd.

6.2. Het LMR-bestand

Bij dit bestand is volgens opgave van de SIG de kwaliteit en de controle zodanig dat codeerfouten zoals tikfouten maximaal circa 0,2% bedragen. Deze foutkans is in principe bij iedere variabele aanwezig. In het volgende overzicht worden die variabelen besproken die een afwijkende foutkans vertonen.

Geslacht. Dit gegeven is medisch van belang en zal daarom zeer nauwkeurig worden ingevuld; des te meer om later de rekening vergoed te krijgen. Wel zou een verschil kunnen optreden met het door de politie opgegeven geslacht omdat men in ziekenhuizen meer geneigd is het door een transseksueel gewenste geslacht te coderen. De omvang van deze verschillen zal naar verwachting gering (< 0,5%) zijn. 'Onbekend' komt (praktisch) niet voor.

Geboortedatum. Ook hiervoor geldt dat juiste invulling in de status van de patiënt belangrijk is voor identificatie en controle. Ook hier zouden fouten leiden tot problemen bij de inning van de rekening. 'Onbekend' komt niet voor.

Datum opname. Doordat deze datum in combinatie met de ontslagdatum gebruikt wordt bij het opmaken van de rekening zullen hooguit fouten van enkele dagen de controle passeren.

Tijdstip opname. Deze wordt door de ziekenhuizen in hele uren naar beneden afgerond. Door drukte bij de ziekenhuisadministratie kan het tijdstip heel goed naar later verschoven zijn, maar zeer onwaarschijnlijk naar voren.

Ziekenhuis(nummer). Uit de aard van de zaak is dit gegeven foutloos. Wèl moet bij de koppeling rekening worden gehouden met de gevolgen van het voortdurende fusieproces tussen ziekenhuizen, om een ondubbelzinnige matching te waarborgen.

Vervoerwijze slachtoffer. Dit gegeven wordt gecodeerd in het vierde cijfer van de E-code, na de punt. Deze codering is op 1 januari 1984 grondig gewijzigd. Bij de proefkoppeling van 1987 - over het jaar 1985 - bleek dat een aanmerkelijk deel van de codeurs nog de oude codering hanteerde (Blokpoel & Polak, 1991). De indruk bestaat dat de kwaliteit van invulling

van de vervoerwijze nog steeds te wensen overlaat. Er komen vrij veel combinaties van dit gegeven voor met het - ook in de E-code opgenomen - type ongeval, die onbestaanbaar zijn. In 1993 waren dat in totaal 615 van de 19.000 (3,2%).
Een voorbeeld is E812.0. Hiervan betekent E812: een " .. verkeersongeval met een motorvoertuig door botsing met een ander motorvoertuig", terwijl .0 betekent dat het slachtoffer een voetganger is. Deze fout kan komen doordat de codeur de oude betekenis van .0 gebruikte: 'Bestuurder van een motorvoertuig (behalve motorfiets of bromfiets)', maar ook door slordigheid. De SIG controleert binnengekomen gegevens niet op onmogelijke E-code combinaties. Wèl zijn zij van mening dat in veel van deze gevallen waarschijnlijk het vierde cijfer (de vervoerswijze) correct is en het type ongeval onjuist.
Ook komt de code voor 'onbekend' veel voor: de vervoerwijze van het slachtoffer was bij 1.760 (9%) onbekend. De ziekenhuizen verschillen ook sterk in het aandeel 'onbekend'. Samenvattend kan van dit gegeven gesteld worden dat foutenpercentages tot enkele tientallen kunnen voorkomen, in samenhang met andere gegevens als de ziekenhuiscode.

De botspartner. Hiermee wordt het object bedoeld waarmee het verkeersslachtoffer in botsing is geweest. Dat kan een vervoermiddel zijn, een voetganger of een obstakel. Bij eenzijdige ongevallen (zoals slippen) is geen sprake van een botspartner. Dit gegeven kan op beperkte wijze worden afgeleid uit de E-code vóór de punt in samenhang met de vervoerwijze zoals die volgt uit het cijfer na de punt. Dit gegeven kent hierdoor dezelfde bezwaren als hierboven genoemd.
De botspartner kan meestal worden onderscheiden in Motorvoertuig (waaronder ook de bromfiets gerekend wordt), Trein, Voetganger en object of obstakel. Van de 19.000 opgenomen verkeersslachtoffers in 1993 was bij 3.459 (18%) alleen bekend dat bij het ongeval een motorvoertuig (waaronder ook brom/snorfiets) betrokken was, zonder duidelijke informatie over de positie van het slachtoffer. (Deze groep overlapt de hierboven genoemde met 1.077, waarbij dus zowel de wijze van vervoer van het slachtoffer als de botspartner onbekend zijn).

Bestuurder of passagier. Dit gegeven wordt alleen gecodeerd voor inzittenden van een personenauto. In 1993 was het aandeel 'onbekend' 1.341 van de 4.901 (27%) slachtoffers die inzittenden waren van een personenauto.

6.3. Het VOR-bestand

Dit bestand is geheel gebaseerd op de gegevens die politiefunctionarissen optekenen als ze van een ongeval kennis hebben genomen. Hoewel de instructies voor het vastleggen van de ongevalsgegevens landelijk uniform zijn (zie de Handleiding AVV/BG, 1995, die regelmatig aangepast wordt) bestaat de sterke indruk dat verschillende politiekorpsen er verschillende interpretaties aan geven. Dat kan zeker het geval zijn bij die korpsen die gebruik maken van geautomatiseerde systemen die voor de politie ontworpen zijn. De bestandsbeheerder (AVV/BG) voert een controle op de ingestuurde registratiegegevens uit waarbij onmogelijke en onwaarschijnlijke combinaties van gegevens worden teruggekoppeld naar de insturende instantie. Er zijn geen interne gegevens beschikbaar over de foutkansen per

variabele. Wel bestaat de overtuiging dat fouten in de belangrijke koppelvariabelen zeer zeldzaam zijn, met uitzondering van het gegeven of een slachtoffer naar een ziekenhuis is vervoerd, aldaar is opgenomen, en in welk ziekenhuis dit gebeurde. In het vervolg worden alleen variabelen genoemd waarover iets bijzonders bekend is.

Geslacht. Dit gegeven zal vanwege het belang voor de identificatie door de politie nauwkeurig ingevuld worden. Wel zal de politie, bij twijfel door bijvoorbeeld transseksualiteit, meer geneigd zijn het officiële geslacht te registreren. Het aandeel onbekend bedroeg in 1992 respectievelijk 1993 en 1994 0,3%, 0,4% en 1,1%.

Datum ongeval. Hier komen geen onbekenden voor. Fouten zijn zeer zeldzaam.

Tijdstip ongeval. De politie kan het tijdstip tot op de minuut registreren. Vaak wordt in de praktijk afgerond op (in volgorde van voorkomen) halve uren, kwartieren, tientallen en vijftallen minuten. Hierbij komen 0,4% onbekenden voor. Omdat de politie dit gegeven uit getuigenverklaringen of schatting moet verkrijgen zullen fouten van de orde van tien à twintig minuten kunnen voorkomen.

Geboortedatum. Ook dit gegeven is voor de taakuitoefening van de politie belangrijk. Het aandeel onbekend was 1,1% in 1992 en 1993, en 1,8% in 1994.

Opname in ziekenhuis. Bij dit gegeven komen vrij veel onbekenden voor: 6,7% in 1992. Voorts is het aannemelijk dat vele procenten ten onrechte als wèl respectievelijk niet opgenomen geregistreerd zijn.

Ziekenhuisnummer. De politie weet meestal niet uit eigen waarneming naar welk ziekenhuis een gewonde vervoerd is, zodat dit gegeven minder nauwkeurig zal zijn. Ook kunnen misverstanden ontstaan door fusies waardoor namen van ziekenhuizen veranderen. Een voorzichtige schatting - op basis van kennis verkregen uit eerdere koppelingen - is dat circa 5% van de geregistreerde ziekenhuiscodes fout is.
De politie geeft in principe niet alleen bij opname, maar ook bij EHBO-behandeling aan naar welk ziekenhuis een gewonde vervoerd is.
Het aandeel met onbekend ziekenhuis onder opgenomen slachtoffers - volgens de politie - wisselt sterk: in 1992 waren dit er 482 (1,0% van de 49.361 slachtoffers), in 1993 waren het 245 slachtoffers (0,5% van de 48.990) en in 1994 waren het 1.007 slachtoffers (2,0% van de 50.513).
Bij de verwerking van de politieformulieren door de codeurs van AVV/BG verschijnt - om de codeertaak te verlichten - na invoeren van de locatie van het ongeval een 'default-ziekenhuis' op het scherm dat het meest in aanmerking komt. Dit zou het maken van fouten echter ook kunnen vergemakkelijken.

Datum van overlijden. Aangenomen wordt dat de politie dit gegeven van het ziekenhuis verneemt, of via justitie (wegens de noodzaak van toestemming voor begraven na een niet-natuurlijke dood). In het belang van een juiste registratie is het belangrijk dat de politie op de hoogte is van de internationale afspraak om gewonden die meer dan dertig dagen na het

ongeval overlijden niet als verkeersdode te registreren. De juistheid van invullen is onbekend.

Vervoerwijze van het slachtoffer. De politie, als deskundige bij uitstek, zal dit gegeven nauwkeurig invullen. Dit gebeurt met een indeling die zeer gedetailleerd is. Ten behoeve van (ongevals)onderzoek wordt deze variabele ingedikt tot een tiental vervoerwijzen. Deze indeling (VVMK) is hier ook gebruikt. De eventuele foutenkans is zoveel kleiner dan die bij de LMR dat hij buiten beschouwing kan worden gelaten.

Botspartner en Bestuurder of passagier. Hiervoor geldt hetzelfde als hierboven gesteld voor de vervoerwijze van het slachtoffer.

6.4. Commentaar

Deze informatie is ontoereikend bevonden om de coëfficiënten van de afstandsfunctie te kunnen schatten. Daarom is in overleg met de opdrachtgever besloten de verkregen informatie aan te vullen door middel van het uitvoeren van een handmatig gestuurde koppeling. Deze is in het volgende hoofdstuk beschreven. De additionele kennis over de foutkansen die is voortgekomen uit deze handmatig gestuurde koppeling wordt besproken in hoofdstuk 8. In deel C is de door de koppeling verkregen kennis over de fouten opgenomen.

7. De handmatig gestuurde koppeling

7.1. Inleiding

Bij de opzet van dit onderzoek werd er van uitgegaan dat de kennis over foutpatronen in beide bestanden in voldoende mate aanwezig zou zijn om een eerste versie van de afstandsfunctie te construeren. Dit bleek echter niet het geval te zijn. De kennis over foutpatronen is daarom verkregen door een handmatig gestuurde koppeling uit te voeren, waarbij steeds één variabele werd onderzocht zonder te eisen dat de waarden gelijk moesten zijn, terwijl alle andere (zoveel mogelijk) exact gelijk moesten zijn. Dit leverde gekoppelde records op waarvan aangenomen mocht worden dat het (over)-grote deel terecht gekoppeld was, zodat de verschillen als fouten geïnterpreteerd konden worden. Door gebruik te maken van deze - niet geplande - tussenstap in het onderzoek werd het mogelijk, zij het met vertraging, toch voldoende gegevens te verzamelen om een automatische koppeling te kunnen uitvoeren. Voorts heeft deze tussenstap veel nuttige informatie voor dit onderzoek opgeleverd.

7.2. Keuze deelverzamelingen

De deelverzamelingen uit de twee bestanden dienden enerzijds zo klein te zijn dat handmatige verwerking in principe mogelijk was, dit omdat alleen op deze wijze directe informatie over aard en omvang van de verschillen in variabelen van te koppelen records verkregen kan worden. Anderzijds moesten de deelbestanden zo groot zijn dat voldoende informatie verkregen zou worden.
Er is voor gekozen om uit beide bestanden alle records te selecteren die een verwijzing naar het Academisch ziekenhuis van de Vrije Universiteit van Amsterdam bevatten. Dit is namelijk het ziekenhuis met het grootste aantal opgenomen verkeersslachtoffers. In het VOR-bestand gaat het in het jaar 1992 om ongeveer 850 slachtoffers, die - volgens de politie - daar opgenomen zijn, of naar dit ziekenhuis vervoerd zijn zonder opgenomen te zijn (dus bijvoorbeeld poliklinisch behandeld), of waarvan het al of niet opgenomen zijn onbekend is. In het LMR-bestand gaat het om ongeveer 480 opnamen.

7.3. De eerste koppeling

De eerste koppeling is handmatig uitgevoerd op de hierboven omschreven bestanden, uit het jaar 1992. Beide bestanden zijn eerst gesorteerd op 'epoch' van opname respectievelijk ongeval. Onder 'epoch' wordt hier verstaan de datum, gecombineerd met het tijdstip. Slachtoffers met dezelfde epoch-waarde zijn gesorteerd op geboortedatum, gevolgd door geslacht. Gehanteerd werden dus de klassieke vier koppelkenmerken, plus het tijdstip.
Vervolgens zijn van beide bestanden de eerste 140 records vergeleken. Het bleek dat behalve 39 koppelingen met afstand nul voor de eis gelijke datum, geboortedatum en geslacht (en ziekenhuis van opname), nog vier koppelingen mogelijk waren door het toestaan van een (gering) verschil in geboortedatum. Bijvoorbeeld 17-07-1962 versus 07-07-1962.

Binnen de koppelingen met afstand nul gold dat het tijdsverschil tussen ongeval-epoch en opname-epoch zelden meer was dan drie uur. Het tijdstip is dus behoorlijk selectief.
Eén koppeling - met afstand nul voor de vier klassieke koppelkenmerken - kwam tot stand met een volgens de LMR overleden persoon, die volgens de VOR niet overleden was. Dit geval is verder afwijkend in die zin dat de E-code in de LMR 928.99 is, dus een ongeval met niet gespecificeerde oorzaak, op een niet gespecificeerde plaats.

7.4. Verdere koppelingen

Op basis van deze resultaten is door de SIG geadviseerd de volgende koppeling op een semi-geautomatiseerde vorm uit te voeren, en wel door beide bestanden te sorteren op:
- ziekenhuisnummer;
- epoch;
- geboortedatum;
- geslacht.

VOR-records zonder ziekenhuisverwijzing worden alleen op de drie laatste sleutels gesorteerd.

De SIG heeft ons geadviseerd de koppelprogrammatuur uit te voeren als EXCEL-applicatie. Dit is een zeer geavanceerd spreadsheet. Van de bij EXCEL mogelijke werkbladfuncties, de mogelijkheid tot het werken met Structured Query Language (SQL), macro's en Dbase-functies, is uitputtend gebruik gemaakt. De beperking van EXCEL tot bestanden met niet meer dan 16.384 records kon eenvoudig omzeild worden door de bestanden op te delen naar epoch, bijvoorbeeld door per kwartaal te analyseren.

Met de hierboven beschreven programmatuur zijn eerst, en wel op enkele testbestanden, de unieke (één op één) koppelingen met afstand nul bepaald en gemerkt met een code die deze afstand aangeeft. Eventuele meerlingen worden op het scherm zichtbaar gemaakt ter beoordeling door de onderzoeker.
Daarna wordt de eis van gelijke geboortedatum verzwakt tot een toegestaan verschil in één deelveld (dag, maand, jaar). Deze worden weer aan de onderzoeker getoond ter beoordeling. Bij goedgekeurde koppeling wordt ook een specifieke code toegevoegd.
Vervolgens vervalt de eis van gelijk ziekenhuisnummer, waarbij ook de VOR-records zonder ziekenhuisverwijzing meedoen (in 1992: 18.622, waaronder 741 ter plaatse en acht later overleden slachtoffers, 652 waarvan de politie een niet bekend ziekenhuis had opgegeven, 16.458 die als niet opgenomen geregistreerd zijn en 763 die onder de code 'ziekenhuis en opname onbekend' vallen). Hierbij wordt weer wel de eis van gelijke geboortedatum gesteld. Bij meerlingen wordt de onderzoeker wederom ingeschakeld.
Afhankelijk van de omvang van de overgebleven bestanden werd de afstandseis verder verzwakt.

De zo verkregen resultaten werden vergeleken met de eerder handmatig uitgevoerde koppeling. De resultaten dienden exact overeen te stemmen. Hierop is de programmatuur gecontroleerd en herzien.

De programmatuur is daarna toegepast op de al eerder met de hand bewerkte bestanden van het VU-ziekenhuis. Daarbij bleek dat van de 850 VOR-records er 203 met afstand nul gekoppeld konden worden, terwijl het toestaan van enig verschil in geboortedatum nog zes additionele koppelingen opleverde.
De politie had maar van 244 van de 850 records opgegeven dat ze inderdaad opgenomen waren; het is interessant dat daarvan 162 direct koppelbaar waren en nog vier met iets grotere afstand: in totaal dus bijna 70%, wat zelfs beter is dan de bijna 60% die landelijk bij de proefkoppeling over 1985 gevonden is. Van de 476 records die volgens de politie niet opgenomen waren konden er acht gekoppeld worden. Van de 104 waarvan het al of niet opgenomen zijn bij de politie onbekend was bleken 23 records direct koppelbaar en nog twee met iets afwijkende geboortedatum. Op deze wijze konden dus in totaal 209 records gekoppeld worden, een verbetering ten opzichte van de 162 + 4 van ruwweg een kwart.

De plausibiliteit van de met afstand groter dan nul gekoppelde records is onderzocht door te kijken naar de mate van overeenstemming van overige variabelen, zoals de wijze van verkeersdeelname en botspartner. Dit heeft tot bovengenoemde extra koppelingen geleid.

Deze resultaten tonen de bruikbaarheid van de programmatuur en de werkwijze aan, terwijl de gevonden koppelpercentages in overeenstemming zijn met de verwachtingen die gebaseerd waren op de resultaten van de proefkoppeling (Blokpoel & Polak, 1991).

In *Bijlage 1* is een uitvoerige beschrijving van de hand van dhr. dr. D.H.M. Frijters van de SIG opgenomen van de gebruikte bestanden en de daarop toegepaste bewerkingen.

8. De afstandsfunctie

8.1. Inleiding

Het nieuwe van de hier beschreven koppeling tussen de bestanden van de VOR en de LMR is dat sprake is van een probabilistische koppeling: niet alleen worden records gekoppeld die voor alle in aanmerking komende variabelen (de koppelvariabelen) gelijk zijn; ook enig verschil wordt getolereerd. Records die de waarde 'onbekend' hebben doen ook mee voor een of meer koppelvariabelen. Er is een gegeneraliseerde afstand gedefinieerd in de door de koppelvariabelen opgespannen ruimte (de koppelruimte) en er wordt niet alleen gekoppeld bij afstand nul, maar ook bij een afstand groter dan nul. De afstand in de koppelruimte tussen de punten die corresponderen met een record uit het VOR-bestand en een record uit het LMR-bestand is zodanig geconstrueerd dat hij een maat is voor de (on)aannemelijkheid dat de records hetzelfde ongevalsslachtoffer betreffen.

8.2. Keuze koppelvariabelen

De ervaring met de handmatig gestuurde koppeling heeft ertoe geleid dat besloten is de volgende variabelen in de afstandsfunctie op te nemen:
- de epoch (datum + tijd) van opname, respectievelijk ongeval;
- de geboortedatum;
- het geslacht;
- het ziekenhuis(nummer);
- de E-code (alleen in het LMR-bestand);
- de variabele ERNSTSL (alleen in het VOR-bestand).

De eerste vier zijn variabelen die in beide bestanden voorkomen. Voorts is uit elk der bestanden een variabele opgenomen die te maken heeft met de aannemelijkheid dat het slachtoffer tot de doelpopulatie behoort, dus een in een ziekenhuis opgenomen verkeersslachtoffer is.

Uit het LMR-bestand is dit de E-code, waaronder behalve (vermoedelijke) verkeersslachtoffers ook zelfmoord en ongevallen met onbekende oorzaak zijn geselecteerd.

Uit het VOR-bestand is de variabele ERNSTSL meegenomen. Deze variabele is door de SWOV aan het VOR-bestand toegevoegd. Zij is geconstrueerd uit alle variabelen die iets zeggen over de ernst van de verwonding, zoals het gegeven of de patiënt is overleden, samen met de tijd die sinds het ongeval is verstreken, of de patiënt is vervoerd naar een ziekenhuis en of hij daar is opgenomen.

Variabele ERNSTSL kent de volgende waarden, waarbij rekening moet worden gehouden met het feit dat van de later overledenen (ERNSTSL 1 t/m 5) de overgrote meerderheid als 'opgenomen' is gecodeerd, maar dat het ook voorkomt dat ze onder de codes 'niet opgenomen' of 'opname onbekend' geregistreerd zijn:

Waarden van de variabele ERNSTSL:
0: ter plaatse overleden (nooit als opgenomen gecodeerd);
1: de zelfde dag overleden;
2: een dag later overleden;
3: 2-5 dagen later overleden;
4: 6-10 dagen later overleden;
5: 11-30 dagen later overleden;
6: opgenomen in een ziekenhuis;
7: vervoerd naar een ziekenhuis, niet opgenomen;
8: vervoerd naar een ziekenhuis, opname onbekend;
9: niet naar een ziekenhuis;
10: alles onbekend.

8.3. De vorm van de afstandsfunctie

Aan iedere koppelvariabele (i) wordt één afstandcoëfficiënt c_i toegevoegd die de afstand aangeeft als twee records alleen voor die variabele sterk verschillende bekende waarden hebben. Sommige variabelen kunnen in meer of mindere mate verschillen. Een voorbeeld is Epoch: als de opname-datum plus tijdstip veel eerder ligt dan die van het ongeval (een negatief Epoch-verschil) is het praktisch onmogelijk dat het om hetzelfde slachtoffer gaat. Bij een positief Epoch-verschil wordt het steeds minder aannemelijk dat het om hetzelfde slachtoffer gaat naarmate dat verschil groter wordt. Om dit soort verschillen in rekening te kunnen brengen wordt een coëfficiënt ϕ_{ik} ingevoerd die de afstand aangeeft als functie van de mate van verschil *k*. Als de variabele in één of beide bestanden onbekend is wordt de coëfficiënt ϕ_i gebruikt om de afstand te bepalen. Beide coëfficiënten kunnen waarden tussen 0 en 1 aannemen. De afstand tussen twee records volgt dan uit:

$$D = \sum_i c_i \, \delta(\alpha_i , \beta_i)$$

met:
D de afstand,
α_i de waarde van variabele i in het LMR-bestand en
β_i de waarde van variabele i in het VOR-bestand,
en $\delta(\alpha_i,\beta_i)$
$= 0$ als $\alpha_i = \beta_i$, beide bekend;
$= \phi_{ik}$ als α_i en β_i verschillen in de mate **k**, beide bekend;
$= \phi_i$ als α_i en/of β_i onbekend;
$= 1$ als $\alpha_i \neq \beta_i$, beide bekend.

Dezelfde vorm wordt gebruikt bij de koppelvariabelen E-code en ERNSTSL, die maar in één van de bestanden voorkomen. Op de lege plaats staat een punt als dummy-variabele.

De coëfficiënten c_i zijn afhankelijk van:
- de foutkansen;
- de resolutie van de variabele;
- de verdeling over de mogelijke waarden.

Deze coëfficiënten worden des te groter gekozen naarmate de variabele meer verschillende waarden kan aannemen, dus selectiever is.

De coëfficiënten ϕ_{ik} zijn afhankelijk van de verdeling van de verschillen tussen de waarden van de variabele *i* bij zeker bij elkaar horende record-paren. Deze verschillen kunnen onbestaanbaar zijn, maar door fouten ontstaan, zoals een negatief Epoch-verschil of verschillend geslacht. Zij kunnen ook onwaarschijnlijk zijn, zodat de afstand de mate van onaannemelijkheid representeert.

De coëfficiënten ϕ_i zijn afhankelijk van de verdeling van de onbekenden over de werkelijke waarden. Deze verdeling is uit de aard van de zaak ook onbekend. Een eerste schatting wordt verkregen door aan te nemen dat ϕ_i gelijk is aan $1 - 1/r_i$, met r_i als de resolutie van de variabele. Hieronder wordt verstaan het aantal waarden dat de variabele kan aannemen. De kleinste resolutie is 2, zoals bij de variabele geslacht; de grootste, met een resolutie van enkele tienduizenden, komt voor bij de geboortedatum. Door deze keuzen is $\delta(\alpha_i,\beta_i)$ te interpreteren als de kans dat de twee willekeurige records verschillende (werkelijke) waarden hebben voor de *i*-de variabele.
Tot nu toe is nog geen eenheid van afstand gedefinieerd. Omdat de uit de afstand af te leiden aannemelijkheid van juiste koppeling het cruciale gegeven is, wordt die gebruikt om de coëfficiënten te normeren. De grens waarboven het juist zijn van een koppeling twijfelachtig wordt, zal gesteld worden op 100. Dit heeft tot gevolg dat de coëfficiënt van een koppel-variabele die bij verschillende (bekende) waarden nooit tot koppeling mag leiden, veel groter dan 100 gekozen moet worden. Een variabele die op zich nooit een koppeling mag verhinderen zal daarentegen een coëfficiënt krijgen die duidelijk onder de 100 ligt.

8.4. De variabelen in de afstandsfunctie

Bij de handmatig gestuurde koppeling van de eerste twee kwartalen van 1993, gevolgd door een analyse van de koppelmogelijkheden, zijn uitgangs-waarden voor de coëfficiënten bepaald. Daarbij zijn achtereenvolgens eerst recordparen met afstand nul gekoppeld en verwijderd, daarna is geëist dat drie van de vier koppelvariabelen die in elk der bestanden voorkomen, de nummers 1 tot en met 4, overeenstemden, met vrijlating van de vierde. Op die wijze zijn groepen gekoppelde records verkregen die met grote waarschijnlijkheid juist gekoppeld zijn. De verschillen in de vrijgelaten variabele zijn dan met grote aannemelijkheid fouten in één der bestanden. De omvang en aard van de verschillen zijn vervolgens gebruikt om de coëfficiënten van de afstandsfunctie te bepalen.

8.4.1. *Epoch-verschil*

Dit is het tijdsverloop tussen het geregistreerde moment van het ongeval en dat van de opname. De politie geeft datum en de tijd in uren en minuten, waarbij vaak sprake is van afronding naar hele kwartieren en veelvouden van vijf en tien minuten. Het opnametijdstip is alleen in hele uren bekend. Een analyse is gedaan over het eerste kwartaal van 1993, waarbij geboorte-datum, ziekenhuis en geslacht exact moesten overeenstemmen. Van de 2.086 gekoppelde record-paren zijn de epoch-verschillen in een tabel

weergegeven, waarbij, afhankelijk van de aantallen, onderverdeeld is in dagen of uren. Omdat het om gegevens over precies één kwartaal ging, zullen verschillen die in naar tijd onbeperkte bestanden voorkomen des te meer buiten de waarneming vallen naarmate ze groter zijn en zijn verschillen van meer dan plus of min negentig dagen onmogelijk. Verschillen van minder dan ongeveer tien dagen worden praktisch volledig waargenomen en daarbuiten wordt een goede indruk verkregen.

Epoch-verschil	Aantal
1-90 dagen negatief	3
0-1 dag negatief	27
0-1 uur positief	189
1-2 uur positief	545
2-3 uur positief	558
3-4 uur positief	419
4-5 uur positief	149
5-6 uur positief	73
6-7 uur positief	23
7-12 uur positief	39
12-24 uur positief	20
1-2 dagen positief	10
2-3 dagen positief	8
3-10 dagen positief	10
10-90 dagen positief	13

Tabel 1. *Epoch-verschillen 1e kwartaal 1993.*

Bij foutloze bestanden kan een negatief epoch-verschil (groter dan circa ½ uur door het afronden) alleen ontstaan doordat twee verschillende gevallen toevallig dezelfde geboortedatum, geslacht en opnameziekenhuis hadden. De grootteorde van het aantal door toeval op deze wijze gekoppelde paren kan als volgt berekend worden:

Geboortedatum, geslacht en ziekenhuis moeten overeenstemmen. We nemen aan dat er geen of verwaarloosbare correlaties bestaan tussen deze variabelen, zodat de totale kans op gelijk zijn het produkt is van de kansen per variabele.
Ook wordt aangenomen dat de geboortedata gelijkmatig verdeeld zijn over de dagen van het jaar, dus kans op verjaardag gelijk is aan 1/365 = 0,00274. De kans dat twee mensen in hetzelfde jaar geboren zijn is te berekenen als een som over de kwadraten van deze kansen per jaar, over alle in aanmerking komende jaren. Daarvoor is de verdeling van de geboortejaren van de verkeersslachtoffers in het VOR-bestand 1993 gebruikt. Het resultaat is een kans van 0,0209 dat twee willekeurige verkeersslachtoffers in hetzelfde kalenderjaar geboren zijn.

De verdeling over de geslachten (ook bij de VOR) is: 0,679 kans op man en; 0,321 kans op vrouw, zodat de kans op een gelijk geslacht gelijk is aan $0{,}679^2 + 0{,}321^2 = 0{,}564$.
Eenzelfde berekening bij de ziekenhuizen (uit het VOR-bestand van 1993) geeft een kans op gelijk ziekenhuis van 0,0105.

Bij deze berekeningen zijn de verschillen in de verdelingen tussen VOR en LMR verwaarloosd, omdat het hier gaat om de orde van grootte van het resultaat.

We vinden zo een kans van:

$$0{,}00274 * 0{,}0209 * 0{,}564 * 0{,}0105 = 0{,}00000034$$

per paar van een willekeurig record uit het LMR-bestand en een uit het VOR-bestand. Omdat er 5.679 LMR-records en 6.700 VOR-records in het eerste kwartaal van 1993 voorkwamen die bekende waarden hadden voor de vier hier relevante koppelvariabelen, waren er dus 5.679 * 6.700 = 38.049.300 van zulke paren. Naar verwachting zouden dus gemiddeld 12,9 van zulke paren door het toeval overeenstemmen, en bij deze koppeling ten onrechte gekoppeld worden. Dit zou neerkomen op ongeveer 1 per week.

De 3 met negatief verschil groter dan 1 dag, kunnen dus geheel op rekening van het toeval geschreven worden. Waar de grens moet worden getrokken bij een positief verschil is wat minder duidelijk. De 27 met een negatief epoch-verschil kleiner dan 1 dag kunnen geen toeval zijn en er moet aangenomen worden dat hier registratiefouten gemaakt zijn, door afronding op hele uren of anders. De grens van niet meer koppelen dient dus te liggen bij een negatief epoch-verschil dat groter is dan 1 dag.
Aan de positieve kant is geen abrupte grens aanwezig. In de periode van meer dan 3 dagen positief zouden een kleine 10 (12,9 gedeeld door 2 omdat hier alleen positief verschil telt) volgens toeval verwacht worden, terwijl 23 gevonden zijn.
Om redenen van beperking van computertijd (zie het hoofdstuk *Koppel-protocol*) zijn de grenzen bij het epoch-verschil in het algoritme verwerkt dat de afstand tussen paren records berekent. Door deze grenzen op -1 dag en +3 dagen te zetten wordt de benodigde rekentijd met een factor van circa 100 verminderd. Naar verwachting zullen hierdoor enkele tientallen record-paren ten onrechte niet gekoppeld worden. Deze krijgen op deze wijze in feite een afstand oneindig, dan wel groter dan iedere andere die kan voor-komen. Om de gedachten te bepalen zetten we de coëfficiënt van deze variabele (c_1) op 1000.
De grote verschillen in aantallen over de range van epoch-verschillen wordt in rekening gebracht door coëfficiënt ϕ_{1k} van de deltafunctie. Deze wordt op nul gezet in het gebied minus ½ uur tot plus 3 uur. Hierin vallen 82% van alle gekoppelde recordparen. Het negatieve half uur is genomen om met afrondingsverschillen rekening te houden. Het gebied van minus 1 dag tot minus ½ uur krijgt een deltafunctie-waarde van 0,04 wat samen met de coëfficiëntwaarde 1.000 een afstand 40 geeft. Van 3 uur positief tot 1 dag positief wordt de deltafunctie op 0,01 gezet en tussen 1 en 3 dagen positief weer op 0,04. Op deze wijze wordt bereikt dat paren met wat grotere epoch-verschillen wat moeilijker koppelbaar zullen zijn, wat de kans op onterechte koppelingen zal verminderen.

8.4.2. *Geboortedatum*

De foutkansen bij deze variabele zijn onderzocht door de gegevens van het eerste kwartaal van 1993 te koppelen met als eis dat het epoch-verschil tussen de 0 en 3 uur moest liggen en geslacht en ziekenhuis gelijk moesten zijn. Dat leverde circa 400 paren op met verschillende geboortedata (gelijke geboortedata komen niet meer voor, want die waren eerder verwijderd). Hiervan verschilde circa 2/3 deel in meer dan twee posities. Administratieve meerlingen kwamen niet voor.
Tot onze verrassing bleken ruim 80 slechts in één cijfer te verschillen.
De kans dat van twee willekeurige personen de geboortedatum in slechts één cijfer verschilt is te becijferen als een som over de (zes) kansen dat van de relevante zes posities van de geboortedatum vijf van de zes gelijk zijn, en de zesde verschilt. (Omdat leeftijden van meer dan 99 jaar uiterst zeldzaam zijn, zijn bij deze berekening alleen de tientallen en de eenheden van het geboortejaar betrokken.) Daarvoor zijn nodig de kansen op gelijk zijn van de zes posities J_3,J_4,M_1,M_2,D_1,D_2 van geboortedata als (19)69 08 29. Deze kansen zijn bij goede benadering 0,2; 0,1; 0,625; 0,125; 0,33 en 0,1. (De kans van 0,2 voor gelijk tiental in het geboortejaar is een gevolg van de gepiekte leeftijdsverdeling van verkeersslachtoffers.) Hieruit volgt een kans van 0,0017 op slechts één verschillend cijfer. Van de 400 paren zou dus naar verwachting hoogstens een enkele zo weinig verschillen. Het is dus zeer aannemelijk dat hier verschrijvingen in de geboortedatum zijn gebeurd en dat het om bij elkaar horende records gaat.
Ook kwamen verwisselingen van twee cijfers vaak voor, die vergelijkbaar zeldzaam zijn. Dit alles wees er op dat bij het registreren van de geboortedatum meer (kleine) fouten gemaakt worden dan eerder werd aangenomen. Om die reden is de deltafunctie voor geboortedata-verschillen van één cijfer op 0,2 gezet en die bij twee verschillende cijfers op 0,5. Dit levert met een coëfficiënt c_2 van 220 een afstand 0 op voor gelijke geboortedata, een van 44 bij één verschillend cijfer en een coëfficiënt van 110 bij twee verschillen. Op deze wijze kan deze laatste groep in het gekoppelde bestand nader onderzocht worden door de grenswaarde te variëren rond de waarde 100. Bij meer verschillen wordt de afstand 220, wat neerkomt op niet koppelbaar.

8.4.3. *Geslacht*

Bij dit deel van de koppeling van het tweede kwartaal van 1993 zijn de records gekoppeld die een epoch-verschil tussen de 0 en de 2 dagen hadden. Dit betekent dat de opname tot hoogstens twee dagen na het ongeval plaatsvond. Ook moest de geboortedatum exact overeenstemmen. Ten slotte moest het om hetzelfde ziekenhuis gaan. We kunnen eenzelfde soort kansberekening uitvoeren als bij het epoch-verschil. Daarvoor moeten we nog weten wat de kans is dat twee willekeurige records een epoch-verschil binnen de genoemde grenzen hebben. Deze kans is bij benadering 2/90 = 0,022. Het bleek dat er geen administratieve meerlingen voorkwamen, zodat altijd uniek gekoppeld kon worden. Volgens het toeval zouden gemiddeld minder dan één geval per kwartaal verwacht worden:

$$0{,}00274 * 0{,}0209 * 0{,}022 * 0{,}0105 = 0{,}0000000132.$$

Vermenigvuldigd met het aantal mogelijke paren geeft dit 0,503. In elf van de 2.600 gekoppelde records bleek het geslacht verschillend te zijn. Deze elf gevallen zullen dus voornamelijk ontstaan zijn door fouten in een der bestanden (waarbij dus terecht gekoppeld is). In ieder geval kan geconcludeerd worden dat de foutkans voor de variabele 'geslacht' hoogstens zeer klein is. De coëfficiënt c_3 is op 90 gezet zodat alleen als er verder bijna niets verschilt, en er geen betere kandidaten zijn, gekoppeld wordt. De coëfficiënt ϕ_3 is op 0,5 gesteld.

8.4.4. *Ziekenhuis*

Om het voorkomen van verschillende ziekenhuizen bij waarschijnlijk terecht gekoppelde paren te onderzoeken, werden dezelfde kwartaalgegevens gekoppeld met als eis dat het epoch-verschil tussen de 0 en 2 dagen moest zijn en geboortedatum en geslacht gelijk. De kans dat een willekeurig paar bij deze exercitie gekoppeld zou worden is dus (bij benadering):

$$0{,}00274 * 0{,}0209 * 0{,}564 * 0{,}022 = 0{,}00000072.$$

Vermenigvuldigd met het aantal mogelijke paren geeft deze kans een verwacht aantal van 27 per kwartaal. Er werden 90 paren gevonden, zodat het merendeel terecht gekoppeld zal zijn en het verschil in ziekenhuis veroorzaakt wordt doordat de politie vaak niet uit eigen waarneming weet in welk ziekenhuis een slachtoffer daadwerkelijk is opgenomen. De coëfficiënt werd daarom op 50 gezet, zodat in het geval van een verschillend ziekenhuis de andere variabelen niet veel meer mogen verschillen, om koppeling bij een afstand kleiner dan 100 nog mogelijk te maken.

8.5. **De coëfficiënten**

Op basis van de verkregen gegevens zijn de volgenden waarden voor de coëfficiënten vastgesteld, waarbij de nummering van § 8.4 is gevolgd:

1. Epoch-verschil

Dit is het tijdsverloop tussen ongeval en opname, waarbij een positief resultaat ontstaat als de opname na het ongeval plaatsvond. Het wordt gegeven in etmalen, dus 3 uur komt overeen met de waarde 0,125.

$c_1 = 1000$;

$\delta(\alpha_1,\beta_1) = \delta(\alpha_1 - \beta_1)$ (α_1 de epoch van opname en β_1 die van het ongeval):

$\delta(\alpha_1 - \beta_1) = 0{,}04$ als $-1 <= (\alpha_1 - \beta_1) < -0{,}021$
$\delta(\alpha_1 - \beta_1) = 0{,}0$ als $-0{,}021 <= (\alpha_1 - \beta_1) <= 0{,}125$
$\delta(\alpha_1 - \beta_1) = 0{,}01$ als $0{,}125 < (\alpha_1 - \beta_1) <= 1$
$\delta(\alpha_1 - \beta_1) = 0{,}04$ als $1 < (\alpha_1 - \beta_1) <= 3$
$\delta(\alpha_1 - \beta_1) = 1{,}0$ in de overige gevallen.

Onbekend komt bij deze variabele niet voor.

2. Geboortedatum

$c_2 = 220.$

Hier wordt het datumveld gesplitst in de 8 posities JJJJMMDD en voor de bekende waarden geldt:

$\delta(\alpha_2,\beta_2) = 0{,}0$ als alle acht posities gelijk zijn
$\delta(\alpha_2,\beta_2) = 0{,}2$ als alle posities op één na gelijk zijn
$\delta(\alpha_2,\beta_2) = 0{,}5$ als alle posities op twee na gelijk zijn
$\delta(\alpha_2,\beta_2) = 1{,}0$ in de overige gevallen.

ϕ_2 is op 0,25 gezet.

3. Geslacht

$c_3 = 90;$

$\delta(\alpha_3,\beta_3) = 0{,}0$ als ze gelijk zijn;
$\delta(\alpha_3,\beta_3) = 1{,}0$ als ze ongelijk zijn.

$\phi_3 = 0{,}5.$

4. Ziekenhuis

$c_4 = 50;$

$\delta(\alpha_4,\beta_4) = 0{,}0$ als ze gelijk zijn;
$\delta(\alpha_4,\beta_4) = 1{,}0$ in de overige gevallen.

Bij de LMR is het ziekenhuis altijd bekend. Als het ziekenhuis bij de VOR onbekend is, wordt toch een nummer gecodeerd, zodat dezelfde afstand gegenereerd wordt als bij verschillend, bekend ziekenhuis. Deze gevallen komen ook tot uiting in variabele 6, ERNSTSL. Dit is het geval bij de ERNSTSL-waarden 0 (ter plaatse overleden), 9 (niet naar ziekenhuis) en 10 (opname en ziekenhuis onbekend). Om implementatie-technische redenen is er voor gekozen in deze gevallen de afstand bij ERNSTSL op 0 te zetten om dubbeltellen te voorkomen.

5. E-code

$c_5 = 100.$

Omdat het VOR-bestand per definitie alleen verkeersslachtoffers betreft en de δ-functie de mate van overeenstemming tussen de bestanden moet uitdrukken, is de δ-functie hier alleen afhankelijk van de variabele E-code uit het LMR-bestand, die (mede) de aannemelijkheid aangeeft of het om een verkeersslachtoffer gaat. De codes waarbij onder de handmatig gestuurde koppeling minder, weinig of in het geheel geen verkeersslachtoffers aangetroffen zijn, krijgen aangepaste bedragen bij de afstanden opgeteld:

$\delta(\alpha_5,.) = 0{,}9$ als E-code = 817.*, 828.*, 958.* of 988.* (geen botsing of zelfmoord(poging));
$\delta(\alpha_5,.) = 0{,}5$ als E-code = 820.* t/m 825.* (geen openbare weg);
$\delta(\alpha_5,.) = 0{,}55$ als E-code = 928.9* (oorzaak ongeval onbekend);
$\delta(\alpha_5,.) = 0{,}0$ in de overige gevallen.

6. ERNSTSL

Hier geldt dat het LMR-bestand per definitie uit personen bestaat die in ziekenhuizen zijn opgenomen, zodat de δ-functie alleen afhankelijk is van de variabele ERNSTSL uit het VOR-bestand, die correspondeert met de mate van aannemelijkheid van het al of niet opgenomen zijn. De waarden zijn in overeenstemming gebracht met de bij de handgestuurde koppeling gevonden percentages gekoppelden naar ERNSTSL. Daarbij is dan aangenomen dat onder de bij variabele 4, ziekenhuis, genoemde omstandigheden de meeste gekoppelden met 'onbekend' of 'verschillend ziekenhuis' terecht gekoppeld zijn.

$c_6 = 50$.

$\delta(.,\beta_6) = 0{,}0$ als ERNSTSL = 0 (ter plaatse overleden);
$\delta(.,\beta_6) = 0{,}7$ als ERNSTSL = 1 (zelfde dag overleden);
$\delta(.,\beta_6) = 0{,}0$ als ERNSTSL = 2, 3, 4, 5 of 6 (later overleden);
$\delta(.,\beta_6) = 0{,}9$ als ERNSTSL = 7 (naar zhs, niet opgenomen);
$\delta(.,\beta_6) = 0{,}7$ als ERNSTSL = 8 (naar zhs, opname onbekend);
$\delta(.,\beta_6) = 0{,}0$ als ERNSTSL = 9 of 10 (niet naar zhs of alles onbekend).

8.6. Afstanden

Door het vermenigvuldigen van de coëfficiënten c_i met de bijbehorende waarden van de δ-functie kunnen de resulterende waarden voor de afstand berekend worden voor de verschillende klassen van de koppelvariabelen:

1. Epoch-verschil

A = 40 als $-1 <= (\alpha_1 - \beta_1) < -0{,}021$
A = 0 als $-0{,}021 <= (\alpha_1 - \beta_1) <= 0{,}125$
A = 10 als $0{,}125 < (\alpha_1 - \beta_1) <= 1$
A = 40 als $1 < (\alpha_1 - \beta_1) <= 3$
A = 999 in de overige gevallen (zie het Koppelprotocol)

2. Geboortedatum

A = 0 als alle 8 posities gelijk zijn
A = 44 als alle posities op één na gelijk zijn
A = 110 als alle posities op twee na gelijk zijn
A = 55 als de geboortedatum onbekend is
A = 220 als ze op meer dan twee posities verschillen

3. Geslacht

A = 0 als ze gelijk zijn
A = 45 als het geslacht onbekend is
A = 90 als ze ongelijk zijn

4. Ziekenhuis

A = 0 als ze gelijk zijn
A = 50 als ze ongelijk zijn
A = 50 als onbekend is welk ziekenhuis het is

5. E-code

A = 90 als E-code = 817.*, 828.*, 958.* of 988.*
A = 50 als E-code = 820.* t/m 825.*
A = 55 als E-code = 928.9*
A = 0 in de overige gevallen

6. ERNSTSL

A = 0 als ERNSTSL = 0, 2, 3, 4, 5, 6, 9 of 10 (bij de waarden 0, 9 en 10 is het ziekenhuis onbekend, zodat toch A = 50)
A = 35 als ERNSTSL = 1 of 8
A = 45 als ERNSTSL = 7.

Bij verschillen in verscheidene koppelvariabelen worden de afstanden opgeteld.

9. Het koppelprotocol

9.1. Inleiding

Uitgegaan wordt van twee (jaar)bestanden, LMR en VOR, respectievelijk bestaande uit circa 26.000 en 50.000 records. In ieder bestand zijn - behalve identificatievariabelen (bijvoorbeeld het rangnummer in het bestand) - koppelvariabelen en controlevariabelen opgenomen. De eerder beschreven afstandsfunctie bepaalt de afstand tussen ieder record-paar (bestaande uit een LMR-record en een VOR-record). In essentie komt de koppeling er op neer dat recordparen die hetzelfde slachtoffer betreffen opgespoord worden door recordparen te zoeken die een kleine onderlinge afstand hebben, terwijl de afstand van ieder record tot andere records duidelijk groter moet zijn; hierdoor is de koppeling voldoende selectief. Dit gebeurt volgens het volgende koppelprotocol:

Stap 1
Beide bestanden worden geprepareerd in een bepaalde volgorde en aan ieder record wordt een rangnummer toegevoegd. (Deze volgorde kan aangepast zijn aan de programmatuur.)

Stap 2
Aan ieder record uit elk der bestanden worden verwijzingen naar een aantal (k) naaste buren uit het andere bestand alsmede de afstanden tot die buren toegevoegd. Hoewel meestal met de dichtstbijzijnde buur (eerste keuze) gekoppeld zal worden (als het paar aan de eisen voldoet) is toch informatie over meer naaste buren (tweede, enzovoort, keuze) nodig. Ten eerste om te kunnen koppelen als de eerste keus bezet is (doordat die een nabijere buur had), en ten tweede om de selectiviteit van de koppeling te kunnen bepalen. Er is gekozen voor twee naaste buren in plaats van drie of meer (zie discussie onder *Beperking rekentijd*).

Stap 3
Uitgaande van het eerste record A_1 (bij volgende keren het eerste nog niet toegewezen record) uit bestand A wordt gekeken of zijn eerste keus (record B_i) uit het andere bestand B ook A_1 als eerste keuze heeft. Dan kan toewijzing volgen: ze zijn elkaars naaste buur. Als record B_i een ander record - A_j - als eerste keus heeft gaan we door, uitgaande van A_j, net zolang tot een paar naaste buren gevonden is. De zo gevonden records worden gemerkt als toegewezen en krijgen twee variabelen toegevoegd, hun onderlinge afstand A en de selectiviteit S van hun toewijzing, die gelijk is aan het kleinste van de twee verschillen tussen eerste en tweede keuze. Als niet met de eerste keuze gekoppeld kan worden (omdat die al gekoppeld is aan een dichterbij staand record), maar met de tweede keuze, wordt S op dezelfde wijze berekend.
Hierna gaan we weer naar het begin van Stap 3 en bij het eerste nog niet toegewezen record begint de procedure van Stap 3 opnieuw. Na voldoende vaak doorlopen van de procedure zijn alle records van het kleinste bestand toegewezen en eindigt de procedure. Bij deze procedure zal het voor kunnen komen dat relatief ver van elkaar afstaande records aan elkaar toegewezen worden. Bij de latere analyse van het gekoppelde bestand kunnen eisen aan

de maximaal toegelaten afstand gesteld worden, omdat de daarvoor benodigde informatie per gekoppeld paar beschikbaar is.

9.2. Potentiële problemen

Aan de hierboven beschreven procedure kleven een aantal potentiële problemen. Van belang is dat de procedure niet in een eindeloze lus terecht kan komen. Dat geval zou zich voordoen als bij een record A als dichtstbijzijnde B gevonden wordt, bij B wordt A^* gevonden, bij A^* wordt B^* gevonden en bij B^* weer A. In zo'n geval stopt de procedure niet. We zullen bewijzen dat bij redelijk geconstrueerde afstandfuncties dit niet kan voorkomen.

Ten tweede moet de procedure voldoen aan de eis dat het koppelresultaat onafhankelijk is van de volgorde van de records in de bestanden.

Van groot praktisch belang is de mogelijkheid dat zeer veel computertijd wordt verspild bij het berekenen van afstanden tussen paren records die veel te veel verschillen. Ook moet vermeden worden dat van veel paren verscheidene malen de afstand berekend wordt.

De oplossing van deze problemen zullen in de volgende paragrafen toegelicht worden.

9.2.1. *Bewijs van de onmogelijkheid van het voorkomen van een eindeloze lus*

De enige eis die aan de afstandsfunctie gesteld wordt is dat de afstand tussen twee records uit de bestanden *A* en *B* reflexief is: de afstand van record A naar record B is gelijk aan die van B naar A. We zullen die afstanden aangeven door $af(A;B) = af(B;A)$. Een eindeloze lus ontstaat als we uitgaande van record A, via een of meer andere records uit *A*, weer bij A terecht komen. Het eenvoudigste geval is boven genoemd. In dat geval zou moeten gelden:

$$af(A;B) > af(A^*;B),$$
$$af(A^*;B) > af(A^*;B^*),$$
$$af(A^*;B^*) > af(A;B^*),$$
$$af(A;B^*) > af(A;B),$$

wat zou neerkomen op $af(A,B) > af(A,B)$, wat niet zo is. Hetzelfde verhaal zou opgaan bij langere lussen, zodat de stelling bewezen is (voor eindige verzamelingen records).
We zien wel dat dit probleem zich kan voordoen bij gelijke afstanden tot de dichtsbijzijnde buur en de op één na dichtsbijzijnde buur, oftewel Selectiviteit 0. De eventuele aanwijzing van *de* dichtstbijzijnde bij keuze uit records met gelijke afstand moet zodanig zijn dat geen eindeloze lus kan ontstaan, bijvoorbeeld door altijd de eerste van de twee naaste buren (met het laagste rangnummer) te nemen. Dit brengt ons op het volgorde-probleem.

9.2.2. *Volgorde-onafhankelijkheid*

Het koppelprotocol (i.c. Stap 3) zoekt recordparen die elkaars - liefst unieke - naaste buren zijn. Als ze inderdaad uniek zijn, dus met Selectiviteit groter dan nul gekoppeld, speelt de volgorde in de bestanden geen rol. Het zal echter ook voorkomen dat er twee of meer even nabije buren bestaan. Dan wordt gekoppeld met diegene van die buren met het laagste rangnummer, dus afhankelijk van de volgorde.
Zo'n koppeling is vanzelfsprekend zeer twijfelachtig, omdat de kans op juiste toewijzing hoogstens 50% (bij twee even nabije buren) of nog minder (bij meer buren) bedraagt. Deze twijfelachtige toewijzingen zijn echter te herkennen omdat hun Selectiviteit gelijk is aan nul. Bij de selectie van vermoedelijk terecht gekoppelde paren zullen die met S = 0 zeker nader onderzocht worden.
Het zal ook voorkomen dat de naaste buur van een record niet beschikbaar is, omdat die al aan een (voor hem nog nabijer) ander record gekoppeld is. Dan is het mogelijk dat toewijzing volgt aan de op een na naaste buur. Omdat (door de keuze voor *twee* naaste buren) niet bekend is of de *derde* naaste buur dezelfde afstand had als de tweede, zullen toewijzingen kunnen ontstaan die volgorde-afhankelijk zijn. Dit probleem zou bij ieder aantal geïnventariseerde naaste buren kunnen optreden. Naar verwachting is het zeer zeldzaam, en de omvang is af te schatten door het (gering gebleken) aantal koppelingen met de op een na naaste buur, op hoogstens enkele per jaar.

9.2.3. *Mogelijkheden tot beperking van de computertijd*

Het is zinloos met afstanden rekening te houden die zo groot zijn dat nooit toewijzing zal volgen. Daarom hoeven bij de 'naaste-buren-berekening' (Stap 2) geen records vergeleken te worden waarvan de data van ongeval en opname meer dan enkele dagen uiteen lopen. Door de beide bestanden te sorteren op epoch hoeft Stap 2 alleen op geschikte records uitgevoerd te worden. Dit levert een reductie van de rekentijd op van ongeveer een factor 100.

Om de selectiviteit te bepalen moet k minstens 2 zijn. Bij een waarde van 2 wordt een record al niet koppelbaar als zijn twee naaste buren aan nog nabijere records gekoppeld zijn. Het lijkt niet zinvol een record aan de op twee na dichtstbijzijnde te koppelen omdat de selectiviteit van deze toewijzing doorgaans slecht zal zijn. Door k de waarde 2 te geven is voldoende bekend over de selectiviteit bij toewijzing: als beide naaste buren even veraf staan is de kans op een juiste toewijzing hoogstens 50%, in het algemeen een te lage waarde. Als de afstanden verschillen geven de twee afstanden voldoende informatie over de selectiviteit. De extra informatie die bij $k = 3$ beschikbaar komt, lijkt marginaal. Uit een oogpunt van besparing van computertijd ligt de keuze $k = 2$ voor de hand.

9.3. **Uitwerking van het koppelprotocol**

Het koppelprotocol is ten behoeve van de SIG opgesteld in 'Pseudocode', een op computertalen als C lijkende taal, die de bedoeling zo ondubbelzinnig mogelijk uitdrukt, en die kan dienen als uitgangspunt voor de programmeur.

Ten eerste krijgt ieder record in ieder bestand een aantal kenmerken toegevoegd, waarvan sommige pas tijdens de koppeling worden bepaald. Het gaat hier om:

- het rangnummer (R_L voor LMR of R_V voor VOR) in het bestand (na de sorteerslag),
- een *boolean* (K) die de waarden *<onbeslist>*, *<gekoppeld>* of *<niet koppelbaar>* kan aannemen,
- twee velden waar *pointers* (P_1 en P_2, wijzend naar het andere bestand) kunnen worden opgeslagen,
- twee velden met *afstanden* (A_1 en A_2) tot de door de pointers aangegeven records in het andere bestand.
- de *selectiviteit* S.

Ook moet nog een waarde gegeven worden voor de afstand vanaf waar niet meer gekoppeld wordt: *Grens_Afstand* = 200, en een die de waarde oneindig representeert: *Max_Afstand* = 999.
Vervolgens worden beide bestanden gesorteerd op achtereenvolgens *epoch*, *ziekenhuisnummer*, *geslacht* en *geboortedatum*, en worden de rangnummers toegekend.

Dan zijn ook de aantallen records bekend in het LMR-bestand en het VOR-bestand, N_L en N_V.

De bestanden worden geïnitialiseerd door alle toegevoegde velden de volgende waarden te geven:

- K := *<onbeslist>*;
- de pointers P worden = 0 gezet;
- de afstanden A worden = *Max_Afstand* gezet;
- de selectiviteit S wordt op nul gezet.

9.4. De procedure Afstandtoekenning

Deze heeft de vorm:

```
While -1 < <Epoch-verschil> < 3 do
For i = 1 to N_L
   for j = 1 to N_V
     do Afstand(i,j)
   next j
next i
```

Alleen als het epoch-verschil klein genoeg is doet routine *Afstand(i,j)* de volgende dingen (hierdoor hoeven maar ongeveer 1/100 van alle combinaties bekeken te worden):

- De routine berekent de afstand A volgens de bovenvermelde formule tussen het i-de record uit het LMR-bestand en het j-de record uit het VOR-bestand.
- De routine doet een update van de twee kortste afstanden (op volgorde!) tot de tot dan toe behandelde records uit het andere bestand. Dit doet hij voor *beide* records!

if $A < A_2$
  if $A < A_1$
   $A_2 := A_1$; $A_1 := A$
  else $A_2 := A$
end.

Ook worden de bijbehorende pointers aangepast.

Na afloop van de hele procedure zijn alle records voorzien van de afstanden tot de twee dichtstbijzijnde records uit het andere bestand, evenals de bijbehorende pointers.
Daarna kan de Boolean berekend worden: in alle records wordt de K op *<niet koppelbaar>* gezet als geldt $A_1 >$ *Grens_Afstand*. Omdat A_1 nooit groter is dan A_2 gebeurt dit als de afstand tot de twee naaste buren groter is dan de grens.

9.5. De procedure Koppel

Na de voorafgaande preparatiefase kan het eigenlijke koppelen beginnen. Dit heeft als kern een routine die, uitgaande van een bepaald record uit een van de bestanden, met behulp van de pointers paren naaste buren zoekt, de selectiviteit berekent en de overige records als niet koppelbaar merkt.

Start:
Begin met het eerste (met (tot dan toe laagste) rangnummer i) record uit het LMR-bestand dat nog als <onbeslist> gemerkt is:

Zoek:
(met deze routine wordt zowel vanuit LMR naar het dichtstbijzijnde VOR-record gezocht als omgekeerd!)
Zoek het record uit het andere bestand waarnaar P_{L1} verwijst.

If K = <onbeslist> (van het andere bestand) dan do:

  if $P_{V1}(P_{L1}) = i$ then <ze zijn elkaars naaste buren> gosub *verwijzen*:
  [de eerste pointer van het record waarnaar de eerste pointer van dit record verwijst gelijk is aan het rangnummer van dit record]

  else verwissel van standpunt en doe hetzelfde (dus gebruik routine *Zoek*) uitgaande van het VOR-record met rangnummer P_{V1}. Als dit niet tot verwijzing leidt, wordt weer van standpunt gewisseld tot verwijzen volgt. Boven is bewezen dat dit altijd in een eindig aantal stappen zal gebeuren.

  else $P_{L1} := P_{L2}$, ga naar *Zoek*: als beide niet <onbeslist> dan wordt het LMR-record als <niet koppelbaar> gemerkt.

Ga naar *Start:*

Verwijzen:
Bij beide records wordt *K* op <*gekoppeld*> gezet, in variabele *A* wordt de afstand opgenomen en *S* wordt berekend als:

$$S = min(A_{L2}\text{-}A_{L1}, A_{V2}\text{-}A_{V1}).$$

Zo zijn alle LMR-records de revue gepasseerd. Het VOR-bestand moet eventueel nog eenmaal doorlopen worden om alle nog als <onbeslist> gemerkte records als <niet koppelbaar> te merken.

9.6. Controle

Het boven omschreven koppelprotocol is door de SIG geprogrammeerd en uitgevoerd op de jaarbestanden van 1993. Daarbij bleek dat het protocol in hoofdzaak juist werkte, in de acceptabele tijd van drie à vier uur per jaarbestand. De resultaten zijn vergeleken met de eerder handmatig verkregen koppelresultaten, waarbij bleek dat enige bijstellingen in de coëfficiënten noodzakelijk waren, en ook enkele 'bugs' in de programmatuur aan het licht traden.

Daarna is de programmatuur getest op twee speciaal geconstrueerde testbestanden: één in LMR-structuur en één in VOR-structuur. In die bestanden kwamen alle kritische combinaties van waarden van de koppelvariabelen voor, zodat de juiste berekening van de afstand en het op de juiste wijze toewijzen van recordparen gecontroleerd kon worden. Ook hierbij bleken nog enkele bugs te bestaan die verbeterd werden.

De controle is tot dusverre minder uitgebreid geweest dan was voorgenomen. De externe steekproeven bleken om redenen van privacybescherming niet uitvoerbaar te zijn. In deel B zijn aan de hand van de resultaatbestanden aanvullende controles uitgevoerd.

Het programma, dat geschreven is door ir. S. Westen van de SIG, is opgenomen als *Bijlage 2*.

10. Resultaten van de koppeling

10.1. Inleiding

Met de hiervoor beschreven vorm en waarden van de coëfficiënten van de afstandsfunctie zijn de bestanden van de jaren 1992 en 1993 in hun geheel gekoppeld. De resultaten zijn intern bij de SIG beschikbaar in de vorm van een LMR-bestand dat additionele informatie bevat die per record aangeeft:

- de afstand A_1 tot het dichtstbijzijnde record uit het VOR-bestand;
- de afstand A_2 tot het op één na dichtstbijzijnde record uit het VOR-bestand;
- pointers P_1 en P_2 naar die twee records;
- een variabele K die 1 is als het record gekoppeld is en 2 als dat niet zo is; en als $K = 1$:
- de afstand A tot het record waarmee gekoppeld is;
- de pointer P tot het record waarmee gekoppeld is;
- de selectiviteit S;
- het Epoch-verschil tussen LMR- en VOR-record;
- de waarden van alle in het VOR-bestand opgenomen variabelen van het VOR-record waarmee gekoppeld is.

Een vergelijkbaar VOR-bestand was in dit stadium van het onderzoek niet beschikbaar. Voor de beoordeling van de het koppelresultaat zijn de benodigde gegevens verkregen door vergelijking met het gehele VOR-bestand. In deel B is ook een VOR-bestand geanalyseerd met bij de koppeling toegevoegde informatie.

Op verzoek van de SWOV zijn ten behoeve van deze rapportage een aantal tabellen uitgedraaid, waarbij nog een variabele (KOPKWAL) is toegevoegd, die de kwaliteit van de koppeling aangeeft.
De grens van 200 waarboven niet gekoppeld diende te worden (zie § 9.3) is niet als zodanig geïmplementeerd. Daardoor zijn een honderdtal records als gekoppeld gemerkt met een afstand groter dan 200. Bij de analyse levert dit geen problemen op omdat ze ver boven iedere acceptabele afstand zitten.
Gerapporteerd worden resultaten van het jaar 1993. De resultaten over 1992 waren geheel in overeenstemming met die van 1993.

10.2. Afstand versus Selectiviteit

Een belangrijke toetssteen van de koppelingsmethode is de mate waarin de gekoppelde paren records onderscheiden kunnen worden naar de aannemelijkheid van de juistheid van hun koppeling.
In het meest ideale geval zou een grote groep gekoppeld moeten zijn met kleine afstand en grote selectiviteit, en de rest òf niet gekoppeld òf gekoppeld met grote afstand en kleine selectiviteit (en dus verworpen wegens zeer onaannemelijk). Met andere woorden: de combinaties (gekoppeld met) kleine afstand en kleine selectiviteit, en grote afstand met grote selectiviteit moeten zo min mogelijk voorkomen. Om dit te beoordelen zijn tabellen uitgedraaid van Afstand tegen Selectiviteit voor alle gekoppelde records. Zoals eerder vermeld zijn records gekoppeld tot een afstand van 200, waarbij de a priori grens van aannemelijke juistheid bij

A = 100 geplaatst is. Omdat A en S elk veel van de waarden tussen 0 en ruim 200 kunnen aannemen zijn dit onhandelbare tabellen. Na een eerste bestudering zijn A en S in de volgende klassen ingedeeld:

Klassen bij Afstand *A*	Klassen bij Selectiviteit *S*
0	0-39
1-40	40-79
41-65	80-119
66-100	120-159
101-130	160+
131-200	
200+	

Tabel 2. *Klassen van Afstand A en Selectiviteit S.*

De afstand 0 betekent: dat alle variabelen bekend moeten zijn, dat het Epoch-verschil tussen de minus een half uur en plus drie uur moet liggen, dat de geboortedatum, het geslacht en het ziekenhuis moeten overeenstemmen, de E-code tot de beperkte standaardrange moet behoren en het slachtoffer volgens de politie in een ziekenhuis opgenomen was en niet dezelfde dag is overleden.
De klasse 1-40 omvat de volgende afstandswaarden: een epoch-verschil tussen 3 uur en 1 etmaal leidt tot een afstand 10; als de persoon dezelfde dag is overleden of de opname is bij politie onbekend, dan leidt dit tot een afstand 35; een negatief epoch-verschil tussen 1 dag en een half uur of een positief epoch-verschil van meer dan 1 dag (maar minder dan drie dagen) leidt tot een afstand 40.
Samen omvatten deze twee klassen de meest selectieve koppelsleutel die bij de proefkoppeling van 1987 is gebruikt.

Voor 1993 zijn in totaal 14.437 van de 25.923 LMR-records gekoppeld. In 1992 waren dat 14.773 van de 27.040 records. Het gaat hier om alle gekoppelde records, van zeer aannemelijk tot volstrekt onaannemelijk. In de volgende tabel zijn de gekoppelde aantallen LMR-records uit 1993 gegeven voor alle combinaties van afstandsklasse en selectiviteitsklasse.

	S=0-39	40-79	80-119	120-159	160+	Totaal
A = 0	23	86	2302	2708	940	6059
1-40	11	246	880	619	108	1864
41-65	76	690	1049	286	17	2118
66-100	193	336	118	8	1	656
101-130	399	180	30	2	0	611
131-200	2819	206	7	1	0	3033
200+	78	18	0	0	0	96
Totaal	3599	1762	4386	3624	1066	14437

Tabel 3. *Afstandklasse tegen Selectiviteitklasse gekoppelde records, 1993.*

We zien dat verreweg de meeste (6.059) records gekoppeld zijn met afstand 0, waarvan de overgrote meerderheid (5.950 of 98%) met selectiviteit 80 of meer.
Bij het beoordelen van de selectiviteit moet bedacht worden dat de coëfficiënten van de afstandsfunctie impliceren dat verschillen in afstand tot circa 40 al kunnen ontstaan door relatief kleine verschillen tussen twee records: een epoch-verschil van meer dan drie uur geeft 10 en als de politie de opname als onzeker opgeeft leidt dit tot een afstand van 35. Daarom moet een selectiviteit tot 40 als (te) klein worden opgevat.

Slechts 86 records (1,4%) hadden een selectiviteit tussen 40 en 79. De 23 met selectiviteit tussen 0 en 39 hadden alle S = 0 en waren dus administratieve meerlingen, waarbij twee (of meer) records in één of beide bestanden dezelfde waarden hadden voor geboortedatum, geslacht en ziekenhuis, terwijl het epoch-verschil gering was, de E-code tot de standaardreeks behoorde en het slachtoffer volgens de politie was overleden en/of opgenomen. In dit stadium zijn ze als slecht gekoppeld opgevat.

De groep met *A* tussen 1 en 40 is nog aanzienlijk (1.864) en ook hier heeft verreweg het grootste deel (1.607 of 86%) een selectiviteit groter dan 79. Het relatief grotere deel met selectiviteit tussen 40 en 79 (246 of 13%) moet gezien worden in het licht van het feit dat in het algemeen de selectiviteit zal afnemen bij toenemende afstand, omdat het record dat de tweede keus is vaak een afstand van 100 tot 140 kan hebben. Dat is bijvoorbeeld het geval als het toeval wil dat volgens de politie in hetzelfde ziekenhuis binnen 0 tot 3 uur eerder iemand met onbekende geboortedatum en geslacht (*A* = 100) is opgenomen. Ook een VOR-record met de combinatie van volgens de politie 'niet opgenomen', geboortedatum één positie verschillend, en Epoch-verschil meer dan 1 dag (*A* = 134) zal relatief vaak kunnen voorkomen. Het aandeel met lage *S* (11) is verwaarloosbaar.

De groep met wat grotere afstand (41 - 65) kent nog 2.118 gekoppelde records, waarvan 1.352 (64%) met selectiviteit groter dan 79. Een relatief nog groter deel dan bij kleinere afstanden, 690 (33%), heeft selectiviteit 40 -79, en nog steeds is het deel met lage *S* klein: 76 (4%).

De groep met *A* tussen 66 en 100 is duidelijk kleiner (656), waarvan slechts 127 (19%) een selectiviteit groter dan 79 heeft. De ook nog selectieve groep met *S* tussen 40 en 79 is de grootste: 336 (51%). Dit is niet verwonderlijk als rekening wordt gehouden met het veel voorkomen van afstanden tussen de 100 en de 140 (en meer) als tweede keuze.

De groep met afstand tussen 101 en 130 is ook klein (611) en heeft in hoofdzaak een geringe selectiviteit: 399 (65%).

Boven een afstand van 130 vinden we weer een aanzienlijk aantal gekoppelde records (3.033 en de ten onrechte als gekoppeld beschouwde groep van 96), met een overwegend slechte selectiviteit.

We zien dat de bovengenoemde combinaties - kleine afstand met kleine selectiviteit en grote afstand met grote selectiviteit - inderdaad weinig voorkomen. Er is een goed onderscheid te maken tussen de volgende gebieden in het *A* - *S* diagram, waaraan de variabele *KOPKWAL* is toegevoegd:

- maximaal aannemelijke juiste koppelingen: $A = 0$ en $S > 79$, KOPKWAL = A;
- zeer aannemelijke juiste koppelingen: $A = 0$ en $S = 40$ t/m 79 en $A = 1$ t/m 40 met $S > 39$, KOPKWAL = B;
- aannemelijke juiste koppelingen: $A = 41$ t/m 65 met $S > 39$, KOPKWAL = C;
- redelijk aannemelijke juiste koppelingen: $A = 66$ t/m 100 met $S > 39$, KOPKWAL = D;
- twijfelachtige koppelingen: $A = 101$ t/m 130 met $S > 39$, KOPKWAL = E;
- praktisch zeker onjuiste koppelingen: alle overige gevallen, KOPKWAL = F.

We zien dat het koppelresultaat in termen van het totale aantal gekoppelde records relatief weinig verandert als de grens in de buurt van $A = 100$ verschoven wordt. Deze grens van aannemelijke koppeling is bewust nagestreefd, maar de kleine aantallen in die buurt zijn een - niet te beïnvloeden, maar wel zeer gewenst - resultaat. Daarmee is aan de eis voldaan dat het koppelresultaat ongevoelig moet zijn voor kleine veranderingen in de afstandsfunctie.

10.3. **Koppelkwaliteit**

10.3.1. *E-code*

Bij deze koppeling is voor het eerst apart aandacht gegeven aan waarden van de E-code waarvan aangenomen werd dat ze (hoofdzakelijk) geen verkeersslachtoffers volgens de internationale definitie betreffen: de groep ongevallen buiten de openbare weg (E-code 820.* t/m 825.*) en de twee E-codes uit de standaardgroep waarbij geen rijdend vervoermiddel betrokken is (E817.* en E828.*). Ook zijn enkele categorieën toegevoegd om te kunnen beoordelen in welke mate daar verkeersslachtoffers voorkomen. Dat zijn de niet gespecificeerde ongevallen (E-code 928.9*), de zelfmoord(pogingen) (E-code 958.* en 988.*) en enkele typen treinongevallen (E-code 801.* en 805.*-807.*).

KOPKWAL	Standaard	Geen rijdend voertuig	Zelfmoord	Niet gespecificeerd	Geen openbare weg	Totaal
A	5.950					5.950
B	1.939					1.939
C	1.514			478	50	2.042
D	330	6	4	115	8	463
E	146	1	2	60	3	212
F	2.944	29	9	791	58	3.831
Gekoppeld	12.823	36	15	1.444	119	14.437
Niet gekoppeld	5.489	848	152	4.719	278	11.486
Totaal	18.312	884	167	6.163	397	25.923

Tabel 4. *Koppelingskwaliteit naar E-code-groep, 1993.*

Deze groepen zijn in *Tabel 4* onderscheiden van de overige E-codes, behalve de treinongevallen; hier bleek het maar om enkele gevallen te gaan. We zien dat voor alle groepen geldt dat in de KOPKWAL-klassen A t/m D een (wisselend) aantal terecht komt, dat kwaliteitsklasse E minimaal gevuld is en F (de onjuiste koppelingen) weer goed gevuld is. Ook dit patroon is overeenkomstig de bedoeling van deze koppelingsprocedure: het zo goed mogelijk kunnen uitmaken waar de grens tussen aannemelijk juiste en aannemelijk onjuiste koppelingen getrokken dient te worden.

10.3.2. *ERNSTSL*

Alle door de politie geregistreerde slachtoffers zijn bij deze koppeling meegenomen; ook die waarbij volgens opgave van de politie geen ziekenhuisopname heeft plaatsgevonden (ERNSTSL 9). In *Tabel 5* is de variabele ERNSTSL uitgezet tegen de koppelingskwaliteit. Omdat de variabele ERNSTSL alleen in het VOR-bestand voorkomt, bevat de tabel alleen de 14.437 gekoppelde records.
Onverwachts bleken toch enkele ter plaatse overleden slachtoffers (ERNSTSL = 0) koppelbaar, ondanks de (door het dan altijd onbekend zijn van het ziekenhuis ontstane) extra afstand van 50. Dit kan niet anders geïnterpreteerd worden dan dat het om ernstig gewonde slachtoffers gaat die wel opgenomen zijn, hoewel de politie aangaf dat ze ter plaatse waren overleden.
Zoals te verwachten was hebben de groepen 1 t/m 6 (1 t/m 5: overleden, meestal opgenomen en 6: volgens de politie opgenomen) het beste koppelingsresultaat.
Ook te verwachten was dat groep 9 (volgens de politie niet opgenomen en dus ook met extra afstand 50) in hoofdzaak niet terecht koppelbaar zou zijn.
Bij alle groepen loopt de grens tussen kwaliteit D en E door relatief kleine aantallen, zodat een kleine verschuiving in de koppelgrens kleine verschuivingen in de gekoppelde aantallen tot gevolg heeft.

ERNSTSL	A	B	C	D	E	F	Totaal
0*			5	2	1	82	90
1		59	8	20		13	100
2	55	24	10			1	90
3	35	10	5	2		5	57
4	25	6	6			1	38
5	27	10	1	2	1	3	44
6	5.808	1.516	1.063	116	51	614	9.168
7			490	123	33	479	1.125
8		314	81	82	6	129	612
9			347	112	116	2.397	2.972
10			26	4	4	107	141
Totaal	5.950	1.939	2.042	463	212	3.831	14.437
* betekeniswaarden van de variabele ERNSTSL, zie § 8.2.							

Tabel 5. *ERNSTSL tegen koppelingskwaliteit, 1993.*

Op basis van deze gegevens, en vooral de zinvolle grens die in de buurt van de beoogde grensafstand 100 blijkt te liggen, zijn in dit stadium de kwaliteitsgroepen A t/m D als (aannemelijk) terechte koppelingen aangegeven en E en F als onterechte koppelingen. In de uitvoerige analyse van de gekoppelde bestanden in deel B zal beoordeeld worden in hoeverre de 463 recordparen uit groep D van voldoende kwaliteit zijn om inderdaad bij het gekoppelde bestand gerekend te worden.

10.4. Vergelijking met de proefkoppeling van 1987

Om een vergelijking met de eerder uitgevoerde koppeling mogelijk te maken, is het nodig rekenschap te geven van de omstandigheid dat bij die koppeling alleen is gewerkt met wat nu afstand kleiner dan 41 heet. Ook werden alleen records gebruikt die bekende waarden hadden voor de koppelvariabelen. We moeten voor de vergelijking de groepen 'Geen openbare weg' en 'Geen rijdend voertuig' toevoegen. In 1987 werden zo 8.120 van de 19.257 LMR-records gekoppeld, wat neerkomt op 42%. Nu zijn 7.889 van de 19.593 met afstand 0 - 40 gekoppeld: dus 40%. Het wat kleinere aandeel gekoppelden moet gezien worden in het licht van de sindsdien van 71% naar 61% aanmerkelijk gedaalde VOR-registratiegraad (= het aantal ziekenhuisgewonden in de VOR-registratie gedeeld door het aantal verkeersslachtoffers in de standaardgroep van de LMR). Er zijn 9.791 (50%) records tot en met afstand 100 gekoppeld, wat een duidelijke verbetering is.

Om vergelijking met de aantallen VOR-records mogelijk te maken zijn in *Tabel 6* de gegevens uit *Tabel 5* wat gecondenseerd en aangevuld met de relevante aantallen uit het VOR-bestand. Daarbij zijn de kwaliteitsklassen A t/m D samengevat onder de noemer 'Goed gekoppeld' en E en F tot 'Slecht gekoppeld'. De meest rechtse kolom bevat de percentages 'Goed gekoppeld' van het totaal in het VOR-bestand.

ERNSTSL	Goed gekoppeld	Slecht gekoppeld	Totaal gekoppeld	In VOR	Percentage goed gekoppeld
0	7	83	90	706	1,0%
1 - 5	305	24	329	546	55,9%
6	8.503	665	9.168	11.562	73,5%
7	613	512	1.125	16.054	3,8%
8	477	135	612	2.183	21,9%
9 - 10	489	2.624	3.113	17.939	2,7%
Totaal	10.394	4.043	14.437	48.990	21,2%

Tabel 6. *Percentages goed gekoppeld naar ERNSTSL, 1993.*

Belangrijk is dat de percentages 'Goed gekoppeld' voor de verschillende klassen van de variabele ERNSTSL overeenkomen met de verwachtingen die gebaseerd waren op de eerdere proefkoppeling.

Van groep 0, de (volgens de politie) ter plaatse overledenen, zijn toch zeven records goed koppelbaar. Aangenomen moet worden dat toch nog reanimatie geprobeerd is.
De groepen 1 t/m 5, overledenen (die volgens de politie nagenoeg allen zijn opgenomen), zijn voor 56% goed koppelbaar. Het wat lagere percentage dan in groep 6 heeft waarschijnlijk ook te maken met het bijzondere karakter van deze groep.
Groep 6, die volgens de politie opgenomen maar niet binnen dertig dagen overleden zijn, geeft het beste koppelresultaat, met 73,5%. Dit cijfer kan vergeleken worden met de 59,5% bij de koppeling van 1987.
Groep 7, die volgens de politie wel naar een (met name genoemd) ziekenhuis vervoerd maar aldaar niet opgenomen is, blijkt inderdaad voor meer dan 96% niet (goed) koppelbaar; echter, door de grote omvang kunnen toch 613 gevallen gekoppeld worden. Eenzelfde beeld treedt op bij groep 9 (waaraan de kleine groep 10 is toegevoegd) die volgens de politie niet naar een ziekenhuis is vervoerd.
Groep 8, die naar een ziekenhuis is vervoerd, met opname onbekend, blijkt nog voor 22% koppelbaar. Tezamen voegen deze groepen een kleine 1.600 records toe aan de opgenomen slachtoffers die de politie geregistreerd heeft. Bij de analyse in deel B zullen deze uitkomsten in detail onderzocht worden.

11. Conclusies

11.1. Algemeen

De verkregen resultaten bij de probabilistische koppeling van de slachtofferbestanden van de VOR en de LMR leiden tot het oordeel dat op deze wijze belangrijk meer records gekoppeld kunnen worden dan bij de eerder gehanteerde koppelingswijze, waarbij alleen exact overeenstemmende records gekoppeld werden. Van deze extra gekoppelde records is het aannemelijk dat het (over)grote deel terecht gekoppeld is, aangezien de in hoofdstuk 8 geschatte aantallen die volgens het toeval gekoppeld zouden zijn, zeer gering zijn. Een variabele is beschikbaar om de gekoppelde records in te delen naar de mate van aannemelijkheid van terechte koppeling.

11.2. De foutencatalogus

Over de omvang en aard van de fouten in de beide bestanden bleek in de loop van het onderzoek onvoldoende kennis voorhanden te zijn. Daarom zijn extra activiteiten verricht om door middel van een handmatig gestuurde koppeling meer inzicht over de fouten te verkrijgen. In een iteratief proces van herhaalde koppelingen is deze kennis in zoverre verkregen dat de vorm en de coëfficiënten van de afstandsfunctie vastgesteld konden worden.
De foutkansen van de variabelen die pas tijdens deel B van dit onderzoek geanalyseerd worden, zullen dan verkregen kunnen worden, terwijl die van de bestaande foutencatalogus verbeterd worden.

11.3. Het koppelprotocol

Ondanks de grote omvang van de bestanden is het mogelijk gebleken de bestanden te koppelen op een wijze die voldeed aan de gestelde eisen, met name het vinden van een duidelijk gemarkeerde grensafstand van aannemelijke koppeling. Ook was het belangrijk dat ieder record uit het ene bestand vergeleken werd met ieder redelijkerwijs in aanmerking komend record uit het andere bestand, terwijl het koppelingsresultaat onafhankelijk moest zijn van de (arbitraire) volgorde waarin vergeleken werd. Door de wijze van koppelen, die getest is met speciaal daarvoor geconstrueerde testbestanden, is de volgorde-onafhankelijkheid van het resultaat gegarandeerd voor wat betreft de 'goed' gekoppelde recordparen. Volgorde-afhankelijkheid is bij 'slecht' gekoppelde records mogelijk, maar onschadelijk.

11.4. Geschiktheid voor ophogen

Door de toegenomen kennis over de samenstelling en de kwaliteit van beide bestanden en de goede kwaliteit en de grote omvang van het gekoppelde bestand is de geschiktheid voor ophogen optimaal.

11.5. Toepasbaarheid van de methode op andere bestanden

De methode is in principe toepasbaar op andere bestanden, mits voldoende kennis over die bestanden voorhanden is. Daarbij gaat het vooral om kennis

over de foutenkansen van de relevante variabelen. Belangrijk is dat in de loop van dit onderzoek gebleken is dat eventueel ontbrekende kennis over foutkansen via een handmatig gestuurde koppeling aangevuld kan worden. Het is dan wel nodig dat er kennis bestaat over de registratieprocessen waaruit in veel gevallen de aannemelijkheid van individuele koppelingen kan worden afgeleid.

11.6. Beleidsrelevantie

De bij dit onderzoek verkregen kennis kan een aanzienlijke bijdrage leveren tot een betere beantwoording van de in het voorwoord genoemde beleidsvragen.
Het gekoppelde bestand kan, door de grotere omvang dan voorheen, gepaard gaande met een hoge kwaliteit, de politiegegevens van verkeersongevallen en -slachtoffers aanvullen met medisch relevante gegevens als verpleegduur en soort verwonding.
Door de mogelijkheid van vergelijking van overeenkomstige variabelen in beide bestanden, kan de kwaliteit van sommige door de politie geregistreerde gegevens beoordeeld worden.
De resultaatbestanden zijn optimaal geschikt voor de beoordeling van de volledigheid van beide registraties en voor de bepaling van ophoogfactoren.

Deel B: Schatting werkelijke omvang en bepaling ophoogfactoren

1. Inleiding

1.1. Terminologie

Een aantal begrippen heeft in dit rapport een speciale betekenis.

Het onderwerp van dit onderzoek wordt gevormd door een bepaalde groep verkeersslachtoffers. Het gaat om slachtoffers van in een bepaalde periode in Nederland gebeurde verkeersongevallen (volgens de internationale definitie), die als gevolg van dat ongeval in een ziekenhuis zijn opgenomen. Deze groep wordt de doelpopulatie genoemd. Binnen de VOR (altijd verkeersslachtoffers) zoeken we naar opgenomen slachtoffers, binnen de LMR (altijd opgenomen) worden verkeersslachtoffers geselecteerd.

Binnen de LMR kennen we de volgende groepen:

E800-E809	*Spoorweg*(ongevallen);
E810-E819	Ongevallen met een *motorvoertuig* (inclusief bromfiets);
E820-E825	Idem *niet* op de *openbare weg*;
E826-E829	Ongevallen met *overige voertuigen*;
E928.9	*Niet gespecificeerde* ongevallen;
E958, E988	*Zelfmoord* (-poging).

De groepen kunnen ook door de cursieve delen - al of niet afgekort - aangeduid worden. De groepen E810-E819 en E826-E829 heten tezamen ook wel de *Standaardgroep*, omdat deze E-codes de tot dusverre gebruikelijke selectie van de doelpopulatie binnen de LMR vormen.

Binnen de VOR kan het begrip *opgenomen* de (na opname) *overleden* slachtoffers soms wel en soms niet omvatten. Dit blijkt uit de context.

Bij de koppeling zijn een deel van de records uit het LMR-bestand en het VOR-bestand aan elkaar toegewezen, kort gezegd *gekoppeld*. De niet-gekoppelde delen van de beide bestanden heten de *restbestanden*.
Binnen de groep gekoppelde records kan een deel onderscheiden worden dat aan hoge eisen van geringe koppelafstand en grote selectiviteit voldoet, dit deel is *goed gekoppeld*. Het overige (gekoppelde) deel is *slecht gekoppeld*. Binnen beide delen komen recordparen voor die bij hetzelfde slachtoffer horen, deze heten *terecht gekoppeld*.
Als een groep gekoppelde records aan de minder strenge eis voldoet dat hij dezelfde slachtoffers omvat in beide bestanden, wordt dat een *correct* gekoppelde groep genoemd. *Goed*, *terecht* of *slecht* gekoppeld slaat dus op individuele gekoppelde records, *correct* altijd op een *groep* gekoppelde records.

Van de correct (en dus a fortiori van de terecht-) gekoppelde records staat vast dat ze tot de doelpopulatie behoren, voor de overige records is dit een van de onderwerpen van dit onderzoek.

1.2. Plaats in het onderzoek

Nadat in deel A van dit onderzoek de feitelijke koppeling van de bestanden van LMR en VOR is uitgevoerd voor de twee jaren 1992 en 1993 bevat dit deel B de analyse van de zo verkregen bestanden.
Deze analyse begon met een eerste controle ter beoordeling van de kwaliteit en de geschiktheid voor het berekenen van ophoogfactoren.
Daarna is een uitgebreide analyse uitgevoerd die tot doel had:

- het beoordelen van de kwaliteit van de gekoppelde records als functie van afstand en selectiviteit;
- het vergelijken van de (goed) gekoppelde deelbestanden met de rest-bestanden, in LMR en VOR;
- het bepalen van de aantallen *terecht* gekoppelden als functie van de verschillende combinaties van afstand en selectiviteit;
- het schatten van het aantal dat tot de doelpopulatie behoort in de doorsnede van beide bestanden;
- het schatten van de aantallen die tot de doelpopulatie in de (rest)-bestanden LMR en VOR behoren;
- het - zo mogelijk - schatten van het aantal dat tot de doelpopulatie behoort dat in geen van beide bestanden voorkomt.

In deze analyse gaat het niet alleen om totalen, maar ook om onder-verdelingen die relevant zijn voor de ophoging, zoals de vervoerswijze.

Ten slotte zijn op basis van de verkregen informatie ophoogfactoren berekend die, uitgaande van beschikbare VOR- en/of LMR-cijfers, de geschatte werkelijke aantallen verkeersslachtoffers uit de doelpopulatie opleveren.

2. Kwaliteitscontrole

Dit hoofdstuk bestaat uit een beschrijving van de voor de analyse beschikbare bestanden, een beschrijving van de uitgevoerde analyses en de resultaten daarvan, gevolgd door conclusies over de kwaliteit van de koppeling en de mogelijkheden tot het berekenen van ophoogfactoren. Op basis hiervan kon worden besloten door te gaan met een uitgebreide analyse die bestond uit een nadere beoordeling van de kwaliteit van het koppelresultaat en een vergelijking van de gekoppelde bestanden met de restbestanden.

2.1. Beschrijving bestanden

Oorspronkelijk was voorzien om de analyse van de gekoppelde bestanden bij de SIG te (doen) plaatsvinden, vanwege het privacy-gevoelige karakter van deze bestanden. Ze bevatten immers tot individuele personen en ziekenhuizen herleidbare gegevens. Het bleek echter op korte termijn niet mogelijk bij de SIG te beschikken over een geschikte werkplek voor de betreffende SWOV-medewerker. Daarom is besloten de analyses zo veel als mogelijk bij de SWOV uit te voeren op 'geanonimiseerde' bestanden. Alleen als de noodzaak zich voordeed te werken met de complete bestanden is dat bij de SIG gebeurd.

2.1.1. *Bij de SIG berustende bestanden*

Deze bestaan - voor de jaren 1992 en 1993 - uit een LMR-bestand en een VOR-bestand, zoals omschreven in deel A, met bij de koppeling gegenereerde informatie, over de overeenkomsten tussen records, uitgedrukt in 'afstand' en de uniekheid, uitgedrukt in 'selectiviteit'.

Het LMR-bestand

Het LMR-bestand bestaat uit het oorspronkelijke LMR-bestand, per record aangevuld met:
- de afstanden A1 en A2 tot de naaste buur en de op één na naaste buur uit het VOR-bestand;
- pointers P1 en P2 naar die twee naaste buren;
- de variabele K met waarde 1 of 2 als het record wel of niet gekoppeld is;

en als K = 1:
- de afstand A tot het VOR-record waarmee gekoppeld is (meestal gelijk aan A1);
- de pointer P tot dat record;
- de selectiviteit S van de koppeling;
- alle variabele-waarden van het gekoppelde VOR-record.

Het VOR-bestand

Het VOR-bestand bestaat uit het door de SWOV aan de SIG ter beschikking gestelde bestand, per record aangevuld met:

- de afstanden A1 en A2 tot de naaste buur en de op één na naaste buur uit het LMR-bestand;
- pointers P1 en P2 naar die twee naaste buren;
- de variabele K met waarde 1 of 2 als het record wel of niet gekoppeld is;

en als K = 1:
- de afstand A tot het LMR-record waarmee gekoppeld is (meestal gelijk aan A1);
- de pointer P tot dat record;
- de selectiviteit S van de koppeling.

2.1.2. *De geanonimiseerde bestanden*

Om de herleidbaarheid van de LMR-records naar individuen of afzonderlijke ziekenhuizen onmogelijk te maken zijn uit het - voor onderzoek bij de SWOV bestemde - LMR-bestand een aantal variabelen verwijderd. Dat zijn alle identificatie-variabelen als patiënt-nummer, opnamenummer enzovoort, de geboortedatum, datum en tijdstip opname en datum en tijdstip ongeval, het ziekenhuis-nummer en de afstanden (en pointers). Wel zijn de leeftijd in jaren, de maand van opname en de afstandklasse geleverd.

Het geanonimiseerde LMR-bestand heeft de volgende variabelen:

GESL	het geslacht volgens LMR;
OPNMND	de maand van opname;
EPOCH	het epochverschil tussen ongeval en opname (in seconden);
OPRED	de opnamereden (1: observatie; 2: diagnose; 3: therapie);
OPURG	de opname-urgentie (0: niet acuut; 1: acuut);
LFTYD	de leeftijd van de patiënt;
WVPLD	het aantal verpleegdagen;
DIAG	de E-code;
AFKLASC	de afstand-klasse (0: 0; 1: 1-40; 2: 41-65; 3: 66-100; 4: 101-130, 5: 131-200, 6: 201+);
S	de selectiviteit;
SELKLASB	de selectiviteit-klasse (1: 0-39; 2: 40-79; 3: 80-119; 4: 120-159; 5: 160+);
K	geeft aan of gekoppeld is;
KOPKWAL	geeft de kwaliteit van koppeling aan (zie Eindrapport Subfase A);

de volgende variabelen komen uit de VOR:

SEXESL	het geslacht;
ERNSTSL	de ernstcode (zie Eindrapport Subfase A);
VVMK	de vervoerswijze;
BOTSP	de botspartner;
FUNC	geeft de functie aan (1: bestuurder; 2: passagier; 3: voetganger);
OPGEN	geeft aan of de gewonde opgenomen is (1: ja; 2: nee, 3: onbekend);
VERVOER	geeft aan of en hoe er vervoerd is naar het ziekenhuis (1: ambulance; 2: eigen gelegenheid; 3: niet vervoerd; 4: onbekend).

Het VOR-bestand is geleverd als boven omschreven, zonder de pointers. In het VOR-bestand bleken een aantal variabelen onbruikbaar geworden. Die zijn opnieuw toegevoegd plus een aantal variabelen die niet aan de SIG waren geleverd. Dit was mogelijk omdat het VOR-bestand nog de variabele KEY_SLA bevatte die het record uniek aanduidt. Het gaat om de variabelen Datum, Tijdstip ongeval en X (deze variabele geeft aan of een onbekend tijdstip ongeval is gecodeerd als 00.00 uur), terwijl de variabelen GEM en PROV, die gemeente en provincie van het ongeval aangeven, zijn toegevoegd.

2.2. De voorlopige analyses

In eerste instantie zijn een aantal beknopte analyses uitgevoerd, voornamelijk gericht op de beoordeling van de kwaliteit van de bestanden maar ook gericht op een eerste schatting van de aandelen ten onrechte gekoppelde respectievelijk niet-gekoppelde records in de verschillende afstandklassen. Van de later uitgevoerde uitgebreide analyse wordt in hoofdstuk 3 verslag gedaan.

2.2.1. *Eerste beoordeling kwaliteit*

De bestanden zijn geleverd in een door het programma SPSS geproduceerd *portable* formaat dat leesbaar is met het bij de SWOV gebruikte programma SAS. Van de bestanden zijn voor een eerste controle dezelfde tabellen uitgedraaid zoals beschreven in deel A, waarna ze onderling vergeleken werden. Daarna is een uitgebreide reeks tabellen uitgedraaid waarmee de volgende controles zijn uitgevoerd.
Voor een eerste controle (die opgevat kan worden als een acceptatie-test) zijn uit alle vier bestanden (LMR en VOR, 1992 en 1993) de verdelingen over een aantal variabelen uitgedraaid naar K. Dit diende vooral ter eerste beoordeling van de aannemelijkheid van juiste werking van de koppel-programmatuur, met inbegrip van de laatste daarin aangebrachte wijzigingen, en om de jaren 1992 en 1993 te vergelijken. Uit de VOR-bestanden zijn beoordeeld: BOTSP, ERNSTSL, FUNC, LFTSL, OPGEN, SEXESL, VERVOER, VVMK en ZIEKHNR. Uit de LMR-bestanden: DIAG, LFTYD, OPNMND, OPURG en WVPLD.
Het bleek dat de jaren zeer sterk overeen kwamen, dat de aantallen met K=1 exact gelijk waren en de verdelingen zoals verwacht werd op basis van de handmatig gestuurde koppeling en de vorige geautomatiseerde koppelingen. Hierbij moest rekening gehouden worden met het feit dat onder K=1 zowel terecht gekoppelde als zeker ten onrechte gekoppeld records kunnen voorkomen. De verdelingen in de gekoppelde en de niet-gekoppelde delen van de VOR-bestanden kwamen in grote lijnen overeen en de niet-gekoppelde delen van de LMR-bestanden bevatten bijvoorbeeld, zoals verwacht, een veel groter aandeel fietsers.
De conclusie was dat de bestanden tot zoverre in orde waren.

Daarna zijn een aantal tabellen uitgedraaid tegen de variabele KOPKWAL die de kwaliteit van de koppeling uitdrukt. (KOPKWAL 1 betekent gekoppeld met afstand 0 en selectiviteit groter dan 79, KOPKWAL 2 betekent A=0 en S= 40-79 of A=1-40 en S >=40 enzovoort. Onder KOPKWAL 6 vallen alle gekoppelde records met selectiviteit kleiner dan 40 en die met A > 100. Zie verder deel A).

Dit is enerzijds gedaan ter verdere controle en voor een eerste beoordeling van de aantallen en aandelen goed gekoppelde records in beide bestanden. Ten tweede kan zo een indruk verkregen worden welke variabelen andere verdelingen hebben in de gekoppelde en de niet-gekoppelde delen van de LMR- en VOR-bestanden.
Voor deze analyse zijn de tabellen beoordeeld met de variabelen Provincie, E-code, OPGEN, OPURG, LFTYD, WVPLD. Ze zijn beoordeeld op plausibiliteit op basis van kennis van de bestanden en de eerdere koppeling. Op deze manier kunnen indicaties verkregen worden over fouten bij de aangeleverde bestanden en de juiste werking van de koppeling. De tabellen gaven geen indicatie over mogelijke problemen.
Interessante resultaten zijn de verdelingen Goed gekoppeld (KOPKWAL 1-4) versus Slecht of niet gekoppeld (KOPKWAL 5-6 en niet gekoppeld) tegen OPGEN, de variabele die binnen de VOR-bestanden aangeeft of het slachtoffer volgens de politie *opgenomen* of *niet opgenomen* was, dan wel dat het opgenomen zijn *onbekend* was.
De resultaten zijn weergegeven in *Tabel 1* en *2*. In het vervolg worden de koppelresultaten uitgebreider geanalyseerd en besproken.

VOR 1992	Goed gekoppeld	Slecht of niet gekoppeld	Totaal	Percentage goed gekoppeld
Opgenomen	8919	3189	12108	73,7%
Niet opgenomen	1122	32804	33926	3,3%
Opname onbekend	679	2648	3327	20,4%
Totaal	10720	38641	49361	21,7%

Tabel 1. *Aantallen goed en slecht of niet gekoppelde records en het percentage goed gekoppeld in het VOR-bestand naar OPGEN, 1992.*

VOR 1993	Goed gekoppeld	Slecht of niet gekoppeld	Totaal	Percentage goed gekoppeld
Opgenomen	8798	3224	12022	73,2%
Niet opgenomen	1085	33015	34100	3,2%
Opname onbekend	511	2357	2868	17,8%
Totaal	10394	38596	48990	21,2%

Tabel 2. *Aantallen goed en slecht of niet gekoppelde records en het percentage goed gekoppeld in het VOR-bestand naar OPGEN, 1993.*

We zien dat de politie in grote lijnen redelijk op de hoogte is van het al of niet opgenomen zijn van verkeersslachtoffers die door hen geregistreerd worden. De *aandelen* die overeenstemmen betreffende het al of niet opgenomen zijn verschillen wel aanmerkelijk: *'opgenomen'* (in 1992) stemt in 73,7% van de gevallen overeen, terwijl *'niet opgenomen'* in 96,7% (100-3,3) overeen komt. Er is in alle drie categorieën een lichte daling in de percentages goed gekoppeld van 1992 naar 1993. Opvallend is een daling in het aandeel 'Opname onbekend' van 1992 naar 1993.

Ook is een analyse gemaakt van de verdeling van de variabele EPOCH die het tijdverschil aangeeft tussen ongeval en opname. Deze kan lopen van minus 24 uur tot plus drie dagen. Deze verdeling bleek sterk te verschillen voor de verschillende koppelingskwaliteiten. Omdat het tijdverschil een rol speelt bij de afstandtoekenning viel dat in principe te verwachten. Maar wel bleek een veel groter dan verwacht aantal records te bestaan met negatief verschil (enkele honderden).
Dit verschijnsel is nader onderzocht met de hulp van een tabel van het gemiddelde van dat tijdverschil, voor de verschillende combinaties van de afstandklasse (AFKLASC) en selectiviteitsklasse (SELKLASB). Daar bleek dat dat gemiddelde bijvoorbeeld binnen de afstandklasse 1-40 sterk varieerde naar selectiviteit, vooral in 1993. (Een negatief tijdverschil van meer dan een half uur levert een bijdrage tot de afstand van 40, evenals een positief verschil van meer dan 24 uur.)
Dit zou kunnen wijzen op onverwacht hoge aandelen niet terecht gekoppelde records in deze groepen, omdat bij in hoofdzaak terecht gekoppelde groepen records geen afhankelijkheid van een variabele als EPOCH van de selectiviteit verwacht wordt. Immers, de selectiviteit zegt iets over de afstand tot de op één na dichtstbijzijnde buur en waarom zou die afstand een rol spelen bij het tijdsverschil tussen terecht gekoppelde records?
De variabele EPOCH komt alleen voor in het geanonimiseerde LMR-bestand, zodat bij de SWOV niet te achterhalen was of dit verschijnsel zich vooral bij bepaalde ziekenhuizen, bepaalde tijden of regio's voordeed.
Omdat het kon wijzen op fouten bij de bestanden of de koppeling, is een nader onderzoek van de groep gekoppelde records met tijdverschil meer negatief dan een half uur gebeurd bij de SIG. Daar bleek dat onder deze groep een zeer hoog aantal records voorkwam met *opname*tijdstip 00 uur. Dit aantal was in 1993 ook groter dan in 1992.
Aangenomen wordt dat het hier gaat om foutief gecodeerde records, waarbij een *onbekend* opnametijdstip is gecodeerd als 0. Bij terechte koppeling leidt dit in verreweg de meeste gevallen inderdaad tot een negatief tijdverschil. Deze records zouden dus bij juiste codering afstand 0 gekregen hebben en een selectiviteit die in veel gevallen 40 hoger zou liggen. Ze zouden op die manier van KOPKWAL 2 naar KOPKWAL 1 verhuizen. Omdat deze beide selecties tot de zeker goed gekoppelde records behoren heeft deze fout geen invloed op de verdere berekeningen.

2.2.2. *Aandeel terecht gekoppelde records*

Voor een onafhankelijke beoordeling van de mate van juistheid van de koppeling, als functie van de verschillende koppelingskwaliteiten zoals uitgedrukt door de variabele KOPKWAL, is gebruik gemaakt van variabelen die in beide bestanden voorkomen maar die geen rol gespeeld hebben bij de afstandtoewijzing, en dus de koppeling. Als deze variabelen overeenstemmen is dat een onafhankelijke bevestiging van de juistheid van toewijzing van twee records, en van de juistheid van de codering.
Tot deze variabelen horen VVMK uit het VOR-bestand en WIJZE uit de LMR. Deze laatste is afgeleid uit de E-code en bestaat uit het deel na de punt. Beide kennen categorieën als Voetganger, Fietser enzovoort. Helaas zijn de definities en de klassen niet exact gelijk zodat de vergelijking pas goed mogelijk is na een vertaalslag.

Gedefinieerd zijn een aantal herkenbaar gelijke combinaties, een aantal dat overeenstemt volgens een oudere codeerlijst, een aandeel 'Niet gespecificeerd' en de rest krijgt de aanduiding 'fout'.
Het blijkt nu dat het aandeel fout in de klasse KOPKWAL 1 t/m 3 ligt tussen 8 en 9%, bij KOPKWAL 4 is het 21% en bij KOPKWAL 5 en 6 rond de 40%. De aandelen fout bij de goede kwaliteitsklassen kunnen bestaan uit werkelijk ten onrechte gekoppelde records, maar kunnen ook ontstaan door fouten bij de registratie.
Omdat de totale mate van overeenstemming bij KOPKWAL 1 zo groot is, is het veel aannemelijker dat het hier om fouten bij de registratie gaat, in dit geval waarschijnlijk bij de ziekenhuizen. Het is ook gebleken dat de als goed gerekende combinaties van vervoerwijzen bij de kwaliteitsklassen 5 en 6 duidelijk vaker voorkomen dan op basis van een (random) niet-terechte koppeling te verwachten was. Naar schatting zitten hier nog circa 30% terecht gekoppelde records onder. In hoofdstuk 3 wordt hier uitvoerig op ingegaan.

2.3. Voorlopige conclusies

Op basis van de tot hier uitgevoerde controles en analyses is het eerdere oordeel bevestigd dat:
- de bij de koppeling gebruikte bestanden zijn zoals beoogd (ze bevatten alle records en alle variabelen volgens de specificatie);
- de koppeling volgens de specificatie is uitgevoerd;
- de afstandsfunctie een maat is voor de aannemelijkheid van de juiste toewijzing van records;
- de goed gekoppelde bestanden een groot aandeel terecht gekoppelde records bevatten;
- de slecht gekoppelde bestanden nog een deel terecht gekoppelde records bevatten;
- de resultaatbestanden een goede basis lijken te vormen voor de berekening van ophoogfactoren;
- er geen aanwijzingen zijn naar onverwachte beperkingen voor de variabelen waarnaar de ophoogfactoren kunnen worden berekend.

2.4. Nadere beoordeling van de kwaliteit van de koppeling

De kwaliteit van het resultaat van de koppeling is op een aantal punten nader beoordeeld. Ten eerste zijn alle uitgedraaide tabellen (meer dan honderd) altijd voor de jaren 1992 en 1993 afzonderlijk bestudeerd op de aanwezigheid van verschillen die boven de statistische ruis uitstegen. Daarbij werd zowel naar verschillen tussen (totale) aantallen als naar verschillen tussen verdelingen (afgezien van totalen) gekeken. Op een enkele uitzondering na, die al besproken is, bleken zulke verschillen niet te bestaan. Dit wijst op een aantal opmerkelijke zaken.
Het blijkt dat deze twee jaren binnen smalle marges gelijk zijn, zowel wat de afzonderlijke bestanden LMR en VOR betreft, als wat hun samenhang betreft. Dit versterkt het vertrouwen in de koppelingsresultaten en in hun toepasbaarheid op gegevens uit latere jaren. In het vervolg zijn daarom alleen gegevens vermeld uit de twee analysejaren gecombineerd.
Ten tweede is de al hierboven genoemde vergelijking van de wijze van verkeersdeelname volgens beide bestanden uitgevoerd voor alle combinaties van afstandklasse, selectiviteitsklasse, ERNSTSL en E-code-

groep. Deze van de koppelprocedure onafhankelijke controle op terecht zijn van de koppeling leidde tot een in dubbel opzicht gunstig resultaat: de koppeling was in de groep met de kleinste afstand zeer goed èn de gegevens bleken in de meeste gevallen door de codeurs juist te zijn ingevuld.
Het bleek namelijk dat in de groep die met afstand 0 gekoppeld was, met een hoge a priori kans op terecht zijn van de koppeling, 8% verschillende vervoerswijzen aangaven. Daarbij zijn de circa 4% gevallen dat vermoedelijk een oude codeerinstructie gevolgd is, niet gerekend.
Onder de 8% met verschillend gecodeerde vervoerwijze komen overigens in hoofdzaak groepen voor die vermoedelijk door een definitieverschil veroorzaakt zijn. Zo lijkt het alsof een stilstaande fietser bij de LMR soms als voetganger gecodeerd wordt, terwijl er ook verwarring lijkt te bestaan tussen fiets en brom(snor)fiets, en de laatste en de motorfiets. Deze codeer-'fouten' worden uitvoerig behandeld in de Foutencatalogus die opgenomen is in deel C van dit rapport.
Ten derde is de combinatie Vervoerswijze van het slachtoffer (binnen de LMR aangegeven met het cijfer na de punt in de E-code) en de *botspartner* (aangegeven door het cijfer voor de punt), het vervoermiddel of het object waarmee gebotst is, dan wel of er van geen botsing sprake was (val, slippen enzovoort) vergeleken met hetzelfde gegeven bij de VOR.
Bij de VOR worden dezelfde categorieën onderscheiden als bij de vervoerswijzen, aangevuld met obstakels. De botspartner kan bij de LMR slechts onderscheiden worden naar *motorvoertuig* (waaronder ook bromfietsen gerekend worden!), *ander voertuig* (fiets, tram, ruiter, paard en/of wagen), voetganger of object. Dit maakt vergelijken minder precies dan bij de vervoerswijzen. Maar ook hier blijkt een grote mate van overeenstemming te bestaan.
Omdat deze vergelijking niet veel toevoegde aan die van de vervoerswijzen wordt er hier verder niet op ingegaan. Wel wijst hij erop dat als in de ziekenhuizen deze gegevens worden ingevuld, en niet als onbekend worden gelaten, ze bijna altijd juist zijn. Het blijft jammer dat er nog in 4% van de gevallen een oude codering is toegepast zodat de verdeling over de vervoerswijzen in de LMR systematisch afwijkt van de bedoelde en blijkbaar in grote lijnen juist waargenomen verdeling.
Hierbij wordt ervan uitgegaan dat de opgave van de politie altijd juist is. Hoewel ook daar fouten gemaakt kunnen worden, pleit voor die aanname dat de politie ter plaatse een onderzoek heeft ingesteld en bij uitstek deskundig is. Wel is het denkbaar dat de politie een meer formele inhoud aan de begrippen zal geven dan via het slachtoffer in de LMR terecht kan komen. Zo zal iemand die een Spartamet berijdt (fiets met bijna onzichtbare hulpmotor die vaak alleen bij tegenwind enzovoort gebruikt wordt), door de politie - zelfs als er gewoon mee gefietst werd - in de categorie brom- en snorfiets gecodeerd worden, terwijl de opgave in het ziekenhuis tot fiets zou kunnen leiden.

2.5. Vergelijking gekoppelde bestanden met de restbestanden

Beide bestanden (LMR en VOR) bestaan uit een (even groot) deel dat gekoppeld is met het andere bestand en een niet gekoppeld deel, dat we het restbestand noemen. Het ligt voor de hand deze delen te vergelijken, maar een vergelijking van de complete bestanden (zoals tot nu toe altijd heeft plaatsgevonden) heeft weinig zin omdat ze bestaan uit delen met geheel verschillende aandelen van de doelpopulatie.

Met name gaat het hier om de deelbestanden die bij deze koppeling zijn meegenomen om te onderzoeken in welke mate er gevallen uit de doelpopulatie in voorkomen, zoals de deelgroep 'niet opgenomen' volgens de politie en de zelfmoordgevallen bij de LMR. Deze verschillende delen zullen daardoor in sterk verschillende mate gekoppeld worden. Omdat deze delen sterk kunnen verschillen in de wijze waarop ze verdeeld zijn over de verschillende relevante variabelen, zoals leeftijd en vervoerswijze, zullen de totale bestanden ook grote verschillen vertonen. Daarom zijn alleen vergelijkingen gemaakt van de verdeling over variabelen binnen die groepen met verschillende aandelen doelpopulatie.
Opvallend was dat die verschillen over het algemeen niet groot bleken te zijn. De door de proefkoppeling van 1987 al bekende grote verschillen bij de leeftijdsverdeling en die naar vervoerswijze tussen gekoppeld bestand en restbestand van de LMR moeten grotendeels aan de verschillende samenstelling van de twee bestanden toegeschreven worden: het gekoppelde bestand bevat - ook bij deze koppeling - veel motorvoertuigongevallen en weinig fietsongevallen (zonder motorvoertuig), terwijl dat bij het LMR-restbestand juist omgekeerd is. Dit hangt samen met de bekende verschillen in leeftijdsverdeling per vervoerswijze.

2.6. Stand van zaken

De nadere beoordeling van de kwaliteit van de koppeling werkt op de volgende manier door in de verdere analyse:

- de gegevens van de twee jaren 1992 en 1993 zijn gecombineerd;
- de verkeerskundige variabelen in de LMR (wijze van deelname en botspartner) blijken (indien niet onbekend) goed overeen te stemmen met de VOR;
- de wijze van verkeersdeelname is geschikt voor het bepalen van het aandeel terecht gekoppelde records;
- de vergelijking van de gekoppelde bestanden met de restbestanden heeft ertoe geleid de ongevallen met motorvoertuigen en de overige ongevallen apart te analyseren.

3. Analysemethode

3.1. Inleiding

Het hoofddoel van dit onderzoek is het berekenen van ophoogfactoren die de gegevens over de doelpopulatie (hierna ook wel te noemen ziekenhuisgewonden), uit één der bestanden VOR of LMR, vertalen naar (geschatte) werkelijke cijfers. Absolute zekerheid over die factoren kan alleen verkregen worden als uit andere bron de werkelijke cijfers over de doelpopulatie bekend zouden zijn. Dit is niet het geval. Gekozen is voor een methode die - uitgaande van de twee officiële registraties die de doelpopulatie omvatten - een schatting mogelijk maakt van die werkelijke cijfers.
De jaarlijkse aantallen geregistreerde ziekenhuisgewonden verschillen sterk tussen LMR en VOR. In het verleden werd aangenomen dat het aantal in de LMR de juiste grootte-orde had en dat het aanmerkelijk lagere VOR-aantal zijn oorzaak hoofdzakelijk vond in het niet ter kennis komen van de politie van een grote groep ongevallen. Een van de doelen van dit onderzoek was na te gaan in hoeverre het VOR-aantal ziekenhuisgewonden inderdaad systematisch te laag is, en of het LMR-aantal ziekenhuisgewonden (als gevolg van een verkeersongeval) inderdaad ruwweg juist is.
In dit hoofdstuk wordt een methode ontwikkeld om de aandelen terecht gekoppelde records te bepalen. De methode is daarna toegepast op de met verschillende koppelingskwaliteit gekoppelde delen van het bestand. De zo verkregen inzichten hebben ertoe geleid de methode definitief toe te passen op een andere indeling van het bestand. Zo kon een betere schatting van de omvang van de doelpopulatie verkregen worden. De resultaten daarvan staan in het volgende hoofdstuk.

3.2. Methode

Om het uiteindelijke doel van de analyse, het berekenen van ophoogfactoren, te bereiken moet een aantal schattingen worden uitgevoerd.

De tot de doelpopulatie behorende aantallen in de afzonderlijke registraties moeten geschat worden. Daartoe moet kennis bestaan over het aantal werkelijke *verkeers*slachtoffers in het LMR-bestand en het aantal werkelijk *opgenomen* personen in het VOR-bestand, waarbij er vanzelfsprekend van wordt uitgegaan dat personen bij de LMR altijd opgenomen en bij de VOR altijd verkeersslachtoffers zijn.

Tevens moet geschat worden hoeveel tot de doelpopulatie behorende slachtoffers in *beide* bestanden voorkomen, in logische terminologie: hoe groot is de *doorsnede* van beide bestanden?Ook moeten de aantallen tot de doelpopulatie behorende slachtoffers worden geschat die slechts in een van beide bestanden voorkomen. Dan is het *totale* aantal in beide bestanden voorkomende slachtoffers (logisch: de *vereniging* van beide bestanden) te berekenen als het totaal van de twee groepen die slechts in één der bestanden voorkomen en de groep die in beide voorkomt.

Om dan het totale aantal ziekenhuisgewonden in Nederland te verkrijgen moet nog een schatting gemaakt worden van het aantal dat in geen van beide registraties terechtkomt. In *Tabel 3* zijn de verschillende deelbestanden aangegeven. De koppeling van LMR- en VOR-bestand vergroot de mogelijkheid bovenstaande schattingen uit te voeren. Als beide bestanden foutloos waren en alleen leden van de doelpopulatie zouden bevatten zou na koppeling alleen de inhoud van de cel *In geen van beide* nog onbekend zijn. De som van de vier cellen geeft dan de totale omvang van de doelpopulatie.

Doelpopulatie	Wel in LMR	Niet in LMR
Wel in VOR	In beide bestanden	Alleen in VOR
Niet in VOR	Alleen in LMR	In geen van beide

Tabel 3. *Verdeling doelpopulatie naar voorkomen in VOR en/of LMR.*

Doordat beide bestanden ook records bevatten die niet tot de doelpopulatie behoren is de situatie ingewikkelder:

Slachtoffers	Wel in LMR	Niet in LMR	Geen doelpopulatie
Wel in VOR	In beide bestanden	Alleen in VOR	Niet opgenomen
Niet in VOR	Alleen in LMR	In geen van beide	
Geen doelpopulatie	Geen verkeersongeval		

Tabel 4. *Indeling van slachtoffers naar voorkomen in VOR en/of LMR en naar wel of niet behoren tot de doelpopulatie.*

Door de aanwezigheid van fouten in de bestanden zijn van de deelbestanden in de tabel alleen benaderingen bekend:

Records	LMR-bestand	
VOR-bestand	Gekoppeld bestand	Restbestand VOR
	Restbestand LMR	

Tabel 5. *Indeling records uit VOR en LMR naar al of niet gekoppeld zijn.*

Hierbij bevat het gekoppelde bestand - bij een geslaagde koppeling - een groot deel van de in het ene zowel als het andere bestand voorkomende slachtoffers, maar er wordt ook een deel gemist dat door fouten een te grote afstand gekregen heeft. Daarnaast bevat het een deel ten onrechte gekoppelde records, die eigenlijk in de restbestanden thuishoren. In de restbestanden zitten - behalve veel terecht niet-gekoppelde, waaronder niet tot de doelpopulatie behorende, records - ook de ten onrechte niet-gekoppelde records.

In het vervolg wordt behandeld hoe vanuit de gegevens volgens *Tabel 5* de cijfers in *Tabel 4* worden bepaald, en daarmee die in *Tabel 3.* Daartoe worden eerst de aantallen terecht gekoppelde records in het gekoppelde deelbestand bepaald (die met grote zekerheid tot de doelpopulatie behoren). Daarna wordt bepaald hoeveel records uit de restbestanden tot de doelpopulatie behoren.

3.3. Werkwijze

De werkwijze is nu als volgt. Uitgegaan wordt van de koppelresultaten zoals geordend volgens *Tabel 5.* Voor de drie cellen van deze tabel wordt het aandeel dat tot de doelpopulatie behoort zo goed mogelijk geschat. Daarbij moet bedacht worden dat in de cel *Gekoppeld bestand* gekoppelde records voorkomen met zeer verschillende koppelkwaliteit en dat een aantal ten onrechte is gekoppeld. Geschat moet worden welk deel terecht gekoppeld is.
Het is niet mogelijk gebleken hierbij gebruik te maken van door steekproeven verkregen gegevens (zoals de oorspronkelijke bedoeling was) en er moest dus een alternatieve methode gevonden worden. Daartoe is de zogenaamde 'footprint-methode' ontwikkeld. Daarbij is gebruik gemaakt van de aanwezigheid van de variabele *wijze van vervoer* in een groot deel van de gekoppelde records, namelijk die uit de 'Standaardgroep', aangevuld met de 'Spoorwegongevallen' en de 'Ongevallen buiten de openbare weg'.
Van deze records wordt ook aangenomen dat ze tot de doelpopulatie behoren.
Van de gekoppelde records die met de hoogste kwaliteit gekoppeld zijn, met afstand 0 en selectiviteit 80 en hoger, wordt uitgegaan. In § 3.5 wordt aannemelijk gemaakt dat het hier gaat om nagenoeg alleen terecht gekoppelde records. Ook is het aannemelijk dat het allemaal ziekenhuisgewonden betreft, omdat de AVV/BG er zeer nauwlettend op toeziet dat alleen *verkeersslachtoffers* in het bestand terecht komen, terwijl door de koppeling met de LMR ook vaststaat dat het om *opgenomen* gewonden gaat.
Daarna worden de met mindere kwaliteit gekoppelde records onder de loep genomen.

Niettegenstaande de bekende problemen met de kwaliteit van invulling van deze variabele bij de LMR, bleek een zodanig sterke samenhang te bestaan tussen de vervoerswijze zoals gecodeerd in de VOR en die in de LMR, dat die gebruikt kon worden om het aandeel terecht gekoppelde records in de verschillende met mindere kwaliteit gekoppelde klassen records te schatten. Vanwege het nieuwe karakter van deze methode wordt in de volgende paragraaf op het principe apart ingegaan.

3.4. De footprint-methode

De footprint-methode wordt aan de hand van een hypothetisch voorbeeld toegelicht. Uitgegaan wordt van een deelverzameling (A) van terecht gekoppelde records en een in beide bestanden voorkomende variabele die hetzelfde gegeven ('kleur' genaamd) op verschillende manier aanduidt, *terwijl die variabele geen rol heeft gespeeld bij de koppeling.*
Met behulp van eigenschappen van deze deelverzameling wordt nu het aandeel terecht gekoppelde records bepaald in deelverzamelingen van met

mindere kwaliteit gekoppelde records. Deze deelverzamelingen worden opgevat als samengesteld uit een deel terecht gekoppelde records en een deel ten onrechte gekoppelde records. Van deze laatste wordt aangenomen dat ze volstrekt willekeurig aan elkaar toegevoegd zijn, in die zin dat de waarde van de variabele kleur volgens de ene registratie stochastisch onafhankelijk is van die in de andere.
Eerst worden dan in een tabel de twee versies van de variabele tegen elkaar uitgezet, horizontaal de LMR en verticaal de VOR:

Records	Rood	Blauw	Onbekend	Totaal
Rood	35	10	5	50
Blauw	5	40	5	50
Totaal	40	50	10	100

Tabel A. *Verdeling terecht gekoppelde records naar 'kleur' in LMR (horizontaal) en VOR (hypothetische gegevens).*

We zien aan de totaal-cijfers in de tabel dat de variabele kleur bij de VOR altijd de waarde Rood of Blauw heeft, en wel even vaak, terwijl bij de LMR 10% onbekenden voorkomen en Blauw wat vaker voorkomt dan Rood. Voorts zien we dat de beide variabelen in 75% van de gevallen overeenstemmen, en in 15% een verschillende kleur aangeven. Dit patroon wordt nu de 'footprint' van terecht gekoppelde records genoemd. Omdat het hier met zekerheid gaat om terecht gekoppelde records worden de 15% op 'buitendiagonaal'-plaatsen geïnterpreteerd als systematische (codeer)fouten en niet als koppelfouten.

Daarnaast bestaat een deelverzameling van met mindere kwaliteit gekoppelde records (B), die er als volgt uitziet:

Records	Rood	Blauw	Onbekend	Totaal
Rood	15	12	3	30
Blauw	9	18	3	30
Totaal	24	30	6	60

Tabel B. *Verdeling minder goed gekoppelde records naar 'kleur' in LMR (horizontaal) en VOR (hypothetische gegevens).*

Met het blote oog is al te zien dat B niet helemaal random is: op de 'diagonaal' (Rood|Rood en Blauw|Blauw) staan wat hogere aantallen dan op de buitendiagonaalplaatsen (Blauw|Rood en Rood|Blauw).

Het is zinvol het gebruik van het begrip random tabel toe te lichten. Onder een random tabel wordt verstaan een tabel zoals zou worden verkregen bij volstrekt willekeurige koppeling van records. Dan zijn de verdelingen van de variabelen kleur uit beide bestanden *stochastisch onafhankelijk.*
Afgezien van statistische fluctuaties zijn de verdeling over de kolommen in

de tabel dan gelijk aan elkaar en aan de totaalkolom. Hetzelfde geldt dan voor de rijen. De (verwachte) waarden in de cellen zijn in zo'n tabel dus direct af te leiden uit de (rand)totalen. Een voorbeeld is de hierna opgenomen *Tabel R'*: een random tabel met dezelfde randtotalen als A.

Records	Rood	Blauw	Onbekend	Totaal
Rood	20	25	5	50
Blauw	20	25	5	50
Totaal	40	50	10	100

Tabel R'. *Random verdeling records naar 'kleur' in LMR (horizontaal) en VOR, met dezelfde randtotalen als Tabel A (hypothetische gegevens).*

Om nu in deelverzameling B het aandeel terecht gekoppelde records te bepalen wordt B opgevat als de som van een random *Tabel R* en een tabel met terecht gekoppelde records T : B = T + R. Uit het feit dat de variabele kleur geen rol heeft gespeeld bij de koppeling volgt dat T dezelfde footprint heeft als A, dus dat geldt T = c A, met c als constante.

Nu wordt de volgende procedure uitgevoerd: een nieuwe *Tabel B'* wordt gevormd door van B een tabel met dezelfde footprint als A af te trekken: B' = B - x A, met x een getal dat aangepast kan worden. Als x = 0 geldt B'= B. Van B' wordt nu getest of hij random is, dat wil zeggen of hij geheel wordt bepaald door zijn randtotalen.

Het testen van de mate van random zijn gebeurt door op basis van de randtotalen van B' een random tabel te construeren en de som S van de absolute waarden van de verschillen te berekenen. Bij een ideale random tabel is deze som nul. Door nu x vanaf nul te laten toenemen zal S in het algemeen afnemen tot een minimum bereikt is. Daarna zal S weer toenemen. *Tabel R* wordt nu gelijk gemaakt aan de *Tabel B'* met minimale S. Bij tabellen uit de praktijk zal het minimum nooit exact nul worden door de aanwezigheid van statistische fluctuaties. Daarom moet dan van de gevonden *Tabel R* nog beoordeeld worden of hij voor het doel van dit onderzoek voldoende weinig van een volledig random tabel verschilt. Dit gebeurt door een combinatie van een beoordeling van de overblijvende - absolute en relatieve - verschillen van R die in S gesommeerd zijn en een CHI-kwadraattest.

Voor *Tabel B* vinden we een minimale waarde S = 0 voor x = 0,4.
De resulterende deeltabellen zijn:

Records	Rood	Blauw	Onbekend	Totaal
Rood	7	2	1	10
Blauw	1	8	1	10
Totaal	8	10	2	20

Tabel T. *Verdeling aandeel terecht gekoppelde records naar 'kleur' in B.*

Records	Rood	Blauw	Onbekend	Totaal
Rood	8	10	2	20
Blauw	8	10	2	20
Totaal	16	20	4	40

Tabel R. *Verdeling aandeel ten onrechte gekoppelde records naar 'kleur' in B.*

De conclusie is dat in *Tabel B* nog eenderde deel, of twintig records voorkomen die terecht gekoppeld waren en veertig die willekeurig aan elkaar toegewezen zijn. Hoewel op deze wijze het aandeel (en het aantal) terecht gekoppelde records in B bepaald is kan niet aangewezen worden welke records dat zijn.

3.5. Vervolg werkwijze

Met behulp van de footprint-methode is het aandeel terecht gekoppelde records bepaald onder de gekoppelde records, voor zover daarin de footprint-variabele (in beide bestanden) beschikbaar was. Dat betreft de Standaardgroep, aangevuld met de Spoorwegongevallen en de Ongevallen buiten de openbare weg.

Anders is het gesteld met de aan de bestanden toegevoegde records die volgens de E-code niet tot de doelpopulatie behoren: de groepen 'Zelfmoord(poging)' en 'Ongeval met ongespecificeerde oorzaak' bij de LMR. Hier is de vervoerswijze bij de LMR vanzelfsprekend niet gecodeerd.
Bij deze groepen met lage a-priori-kans op behoren tot de doelpopulatie zijn de *goed* gekoppelde records als *terecht* gekoppeld beschouwd. Bij *goed* wordt dan gedacht aan een afstand tot 100 en een selectiviteit van 40 of meer. Daardoor is enerzijds een aantal te veel meegerekend (doordat onder de goed, maar met afstand groter dan 0, gekoppelde records ook ten onrechte gekoppelde records kunnen voorkomen), maar anderzijds een aantal te weinig, omdat onder de slecht gekoppelde records ook terecht gekoppelde zullen voorkomen.
Door het ontbreken van de vervoerswijze kan de footprint-methode hier niet toegepast worden. Uit de aantallen goed/slecht/ niet gekoppeld kan afgeleid worden dat de gemaakte fout procentueel niet groot is.

Bij alle groepen records zijn de aantallen terecht gekoppelde records bepaald en opgenomen in *Tabel 11.*

Voor het bepalen van het aantal terecht gekoppelde records in elke kwaliteitsklasse is de footprint gebruikt van de groep met de hoogste koppelkwaliteit. Ten behoeve van de definitieve schatting van de omvang van de doelpopulatie is een nieuwe footprint bepaald, van een grotere groep records. Bij de analyse bleek namelijk dat de footprint van met afstand tot 100 maar met kleine selectiviteit gekoppelde records zoveel leek op die van de best gekoppelde records dat besloten is ze samen te nemen. Bij kleine selectiviteit, en zeker bij S=0, is het niet meer zeker of twee ononderscheidbare (of moeilijk onderscheidbare) records (voorkomend in beide bestanden) aan de juiste gekoppeld zijn of dat ze verwisseld zijn. Toch zijn ze als groep juist gekoppeld. Omdat het begrip *terecht* gereserveerd is voor

individueel juiste koppeling wordt van groepen als deze gezegd dat ze *correct* gekoppeld zijn. Daarom is besloten de footprint-methode voor het definitieve resultaat te herhalen, uitgaande van een gewijzigde definitie van goed gekoppeld, namelijk gekoppeld met een afstand tot 65, zonder eis aan de selectiviteit. De resultaten hiervan en de schatting van de omvang van de doelpopulatie in de restbestanden staan in hoofdstuk 4.
Uit de analyse was ook kennis verkregen over het koppelresultaat van deelverzamelingen. Sommige E-code-groepen hadden een dusdanig laag aandeel goed gekoppelde records, terwijl een nadere bestudering van het codeboek uitwees dat het niet of praktisch niet om verkeersongevallen ging, dat besloten is ze uit de Standaardgroep te halen. Deze groepen zijn gevoegd bij de groepen met lage a priori kans op behoren tot de doelpopulatie (Ongespecificeerde ongevallen en zelfmoord(pogingen)).
Van deze groepen zijn alleen de goed gekoppelde records als behorend tot de doelpopulatie gerekend.

Op deze wijze was dus een schatting bepaald van de in het gekoppelde bestand voorkomende records die tot de doelpopulatie behoren, de *doorsnede*. Het resultaat is weergegeven in *Tabel 13*.

Daarna moesten nog de aantallen behorend tot de doelpopulatie geschat worden in de twee restbestanden. Daartoe zijn de restbestanden onderverdeeld in groepen met verschillende a priori kans op behoren tot de doelpopulatie. Zoals bij alle groepen gold voor de groepen met lage a priori kans dat de aantallen gekoppelde records bij kleine afstand (relatief) hoog waren, maar bij toenemende afstand kleiner werden om bij nog grotere afstand weer toe te nemen. Dit is in overeenstemming met de aanname dat voor deze groepen in de restbestanden weinig ziekenhuisgewonden voorkomen.
Omdat daar zonder nader specifiek onderzoek geen verdere kennis over beschikbaar is wordt nu van de groepen met (zeer) lage a priori kans aangenomen dat de niet correct gekoppelde records niet tot de doelpopulatie behoren. Anders ligt het bij de groepen met hoge a priori kans. Op basis van de koppelresultaten en kennis van de registratieprocessen is de aannemelijkheid onderzocht en geschat dat het om tot de doelpopulatie behorende records gaat die in het andere bestand ontbreken.

Zo zijn dan van drie van de vier cellen uit *Tabel 3* de omvang en inhoud bepaald. De vierde cel is geschat door aan te nemen dat de kans voor een ziekenhuisgewonde om in het LMR-bestand te komen onafhankelijk is van de kans om in het VOR-bestand te komen.

3.6. Bepaling aandeel terecht gekoppelde records

Ten eerste moet vastgesteld worden welk deel van de gekoppelde records inderdaad hetzelfde slachtoffer betreft, dus terecht gekoppeld is. Zoals eerder opgemerkt kon dit aandeel niet - zoals oorspronkelijk was voorzien - door diepgaande analyse van steekproeven uit de records vastgesteld worden. Het bestaan van een grote groep met zeer grote zekerheid terecht gekoppelde records bleek een alternatieve methode op te leveren.
De gekoppelde records in de verschillende afstandklassen en selectiviteitsklassen voldoen aan een groot aantal voorwaarden. Zo moet in de laagste (beste) afstandklasse (A=0) het tijdstip van opname niet meer dan drie uur

later vallen dan dat van het ongeval, en niet meer dan een half uur eerder. Ook mag de geboortedatum niet verschillen, het geslacht moet gelijk zijn en het ziekenhuis moet overeenstemmen. De politie moet aangegeven hebben dat het slachtoffer niet ter plaatse of dezelfde dag overleden is, en in een ziekenhuis is opgenomen. De in het ziekenhuis geregistreerde E-code moet aangeven dat het om een verkeersongeval gaat, maar met uitsluiting van de codes voor 'In- en uitstappen' (817), die voor 'Ongevallen met motorvoertuigen buiten de openbare weg' (820-825) en 'Ongevallen met een bereden dier' (828).
Ondanks deze beperkingen vormt deze groep verreweg de grootste onder de groepen met verschillend zware eisen aan de afstand. In de verschillende selectiviteitsklassen geldt ook nog de eis dat de beide op één na naaste buren een bepaalde minimum afstand verder weg moesten staan dan de afstand tussen de gekoppelde records.
Deze eisen zijn zo restrictief dat in de groep KOPKWAL 1 (A = 0 en S > 79) het aandeel ten onrechte gekoppelde records zeer klein moet zijn. Dit volgt uit de volgende redenering.

Ten onrechte gekoppelde records kunnen op twee manieren ontstaan: ten eerste door het toeval: twee foutloze records lijken op bovenvermelde manier sterk op elkaar zonder hetzelfde slachtoffer te betreffen, terwijl van ieder record de werkelijke tegenhanger in het andere bestand ontbreekt. Bij aanwezigheid van slechts één van die tegenhangers zou immers al met S=0 gekoppeld zijn! Deze manier wordt zeer uitzonderlijk geacht, omdat de kans van optreden zo klein is dat hooguit enkele gevallen per jaar kunnen optreden.
Een tweede manier om in de groep met KOPKWAL 1 ten onrechte gekoppelde records te krijgen is door fouten bij de codering. Twee records met werkelijke afstand groter dan 0 krijgen toch afstand 0 door een (maar precies de juiste) fout in een van beide records. Ook hier moet voor terechtkomen in deze groep dan weer gelden dat het record van de goede tegenhanger ontbreekt (want anders was S=0) of ook (in voldoende mate) fout is. Ook dit moet zeer uitzonderlijk geacht worden.
Een belangrijk gegeven bij de schatting van de kansen op deze manieren is het feitelijk optreden van administratieve meerlingen in de beide bestanden afzonderlijk. Dat is uitermate zeldzaam, enkele tientallen in beide bestanden. Hieronder vallen beide hierboven genoemde manieren van het optreden van 'identieke' records.
Deze administratieve meerlingen komen - indien gekoppeld - vanzelf terecht in de klasse met S=0. Het is dus zeer aannemelijk dat het aantal onterecht gekoppelde records in deze groep nog aanmerkelijk kleiner is dan het aantal administratieve meerlingen, dus hooguit enkele op de vele duizenden in de groep met KOPKWAL 1.
Het bestaan van een - grote - groep met praktisch uitsluitend terechte koppelingen geeft dus een welkome mogelijkheid om de aandelen terecht gekoppelde records in de overige afstands- en selectiviteitsklassen te schatten. Zoals in hoofdstuk 2 beschreven beschikken we namelijk over een paar selectieve variabelen die in hoofdzaak dezelfde betekenis hebben in de twee bestanden en die bij de koppeling geen rol gespeeld hebben. Bij de proefkoppeling over 1985 is gebleken dat de codering van de vervoerswijze in de LMR anders ingedeeld is dan bij de VOR, dat hij minder nauwkeurig werd ingevuld en dat in een aantal ziekenhuizen nog een geheel andere, van voor 1984 daterende, codelijst gehanteerd werd.

Omdat onbekend was in hoeverre deze onvolkomenheden nu nog bestaan is bij deze koppeling afgezien van het betrekken van deze variabelen in de afstandsfunctie. Daardoor geeft vergelijking van deze variabelen uiterst gevoelige en onafhankelijke informatie, zowel over de huidige omvang van deze onvolkomenheden als over de aannemelijkheid van terechte koppeling van groepen gekoppelde records.

3.6.1. *Ongevallen met motorvoertuigen*

In *Tabel 6* zijn deze variabelen tegen elkaar uitgezet, voor de groep E-codes die ongevallen met motorvoertuigen omvatten (810-829), in de hoogste kwaliteitsklasse KOPKWAL 1.

VOR	LMR-codering								
VVM	Voet	Fiets	Brom	Motor	Auto	Vr/B	Overig	Onbek.	Totaal
Voet	849	67	17	1	46	31	1	91	1103
Fiets	311	1544	13	2	119	15	4	152	2160
Brom	54	89	1660	57	28	5	48	66	2007
Motor	8	3	67	811	13	1	5	49	957
Auto	153	50	13	7	4053	17	14	663	4970
Vracht	3				23	23		6	55
Bus						11			11
Rail							2		2
Overig	5	3	3	1	8		9	10	39
Totaal	1383	1756	1773	879	4290	103	83	1037	11304

Tabel 6. *Vervoerswijze (VVM) in de VOR (verticaal) tegen die in de LMR, bij ongevallen met motorvoertuigen in 1992 en 1993, KOPKWAL 1.*

Om de inzichtelijkheid te verhogen zijn niet alle afzonderlijke LMR-codes opgenomen; met name die voor 'Bestuurder personenauto', 'Passagier personenauto' en 'Inzittende Niet Nader Omschreven Personenauto' zijn samengevoegd onder 'Auto'. Ook zijn de 'Bestelauto's' bij de VOR samengevoegd met de '(Personen)Auto', omdat die categorie bij de LMR niet voorkomt en ze verreweg het meest als 'Auto' blijken te zijn gecodeerd bij de LMR. Hier, en in alle volgende tabellen, zijn nullen weggelaten.

Opvallend is het zeer grote aantal dat op de 'diagonaal' ligt, dus met overeenstemmende vervoerswijze. Circa 10% wordt bij de LMR als onbekend ('Niet gespecificeerde vervoerswijze', hier 'Onbek.') gecodeerd, redelijk in verhouding tot de (rand)totalen volgens de VOR-indeling. Toch vinden we op sommige plaatsen vrij grote aantallen buiten de diagonaal. Sommige daarvan zijn te begrijpen als aangenomen wordt dat vier à vijf procent van de LMR-codeurs nog met de oude codelijst werkt.
Een opvallend voorbeeld daarvan vormen de 153 in de cel Auto|Voet en de 50 in Auto|Fiets. De oude LMR-code voor Autobestuurder is na 1984 in gebruik voor Voetganger en die voor Autopassagier is nu Fiets. Ook de 67 in de cel Mot|Bro, de 119 in Fiets|Auto, de 31 in Voet|Vr/B en de 48 in

Bro|Ove zijn zo te verklaren. Volgens opgave van de SIG is het hun bekend dat plaatselijk nog de oude codering wordt gebruikt.
De grootste resterende buitendiagonaalcel, 311 in Fiets|Voet is niet op deze wijze te verklaren. Een nauwkeurige lezing van de codeerinstructies voor de LMR geeft echter aanleiding voor de veronderstelling dat het hier zou kunnen gaan om tijdens het ongeval *stilstaande* fietsers die door de LMR als voetgangers zijn opgevat. Ook de SIG acht dit een aannemelijke verklaring.
De overblijvende buitendiagonaalcellen vormen in totaal circa vijf procent. Aannemelijk is dat het hier in hoofdzaak gaat om terecht gekoppelde records, waarbij min of meer random fouten gemaakt zijn bij de codering, mogelijk door verwarring tussen op elkaar lijkende vervoerswijzen.

Onder deze aannamen kan de verdeling over de cellen van *Tabel 6* (de 'footprint') gebruikt worden om te onderzoeken welk deel van vergelijkbare tabellen met mindere koppelingkwaliteit uit terecht gekoppelde records bestaat. Vergelijkbare tabellen voor KOPKWAL 2 en 3 hebben dezelfde footprint als *Tabel 6*, maar met 3.453 respectievelijk 2.907 gekoppelde paren. Bij KOPKWAL 4, met 509 records, is het aandeel fout gestegen tot een kleine 20%. KOPKWAL 5 heeft met 197 records een foutpercentage van circa 40.

VOR	LMR-codering								
VVM	Voet	Fiets	Brom	Motor	Auto	Vr/B	Overig	Onbek.	Totaal
Voet	90	26	28	3	42	8	3	21	221
Fiets	89	135	120	59	177	19	10	65	674
Brom	54	73	331	64	144	16	17	54	753
Motor	20	11	34	75	42	3	1	26	212
Auto	149	122	217	123	651	33	17	195	1507
Vracht	3	5	6	4	9	3		2	32
Bus	3	1	1		4	1		1	11
Rail	1				3		1	1	6
Overig	2	2		2	2			1	9
Totaal	411	375	737	330	1074	83	49	366	3425

Tabel 7. *Vervoerswijze in de VOR (verticaal) tegen die in de LMR, bij ongevallen met motorvoertuigen in 1992 en 1993, KOPKWAL 6.*

In *Tabel 7* staan de 3.425 records die met KOPKWAL 6 gekoppeld zijn (waaronder een grote groep die met zeer lage selectiviteit gekoppeld is), met dezelfde beperking tot de groep motorvoertuigongevallen als die in *Tabel 6*. We zien dat hier sprake is van een grote achtergrond van random gekoppelde records, met wat grotere aantallen op de diagonaal.
Tabel 7 wordt nu opgevat als de som van twee tabellen: een deel met dezelfde footprint als *Tabel 6* (waarbij aangenomen wordt dat hij correspondeert met de groep terecht gekoppelden) en de rest een random tabel (die onterecht gekoppelde records bevat). Daarbij wordt onder een random tabel verstaan een waarbij iedere cel in dezelfde verhouding staat tot zijn

verticale of horizontale buren als de bijbehorende randtotalen. Het aantal terecht gekoppelde records in *Tabel 7* is nu geschat door er een zo groot deel van *Tabel 6* af te trekken dat de resulterende tabel zo min mogelijk van een pure random tabel verschilde.
De kleinste afwijking van een random tabel werd bereikt na het aftrekken van een tabel met de footprint van *Tabel 6*, met 919 records. De overblijvende tabel had per cel geen grotere verschillen met een op basis van dezelfde randtotalen geconstrueerde pure random tabel dan circa 40.
De CHI-kwadraat was weliswaar significant, maar bij de grote aantallen is een klein systematisch verschil daarvoor al voldoende. Aannemelijk is dus dat de groep KOPKWAL 6 toch nog ruim 900 terecht gekoppelde records bevat en de rest onterecht gekoppelde.

Eenzelfde werkwijze bij de andere klassen levert op dat de klasse 2 en 3 een verwaarloosbaar aandeel random kennen, dat klasse 4 een aandeel van 73% (373 van de 509) terecht gekoppelde records heeft en klasse 5 ruim 30% (62 van de 197).

3.6.2. *Ongevallen met overige voertuigen*

Hierbij speelt de codewijziging bij de LMR van 1984 geen grote rol, want de oude codering van de vervoerswijze bij deze groep ongevallen verschilt alleen op ondergeschikte punten van de huidige. Ook hier is het aantal op de diagonaal groot, maar zeer eenzijdig geconcentreerd op de fietsers. Bij deze groep ongevallen komen 'illegale' vervoerswijzen voor, die niet zouden mogen bestaan bij dit type ongeval: bromfietsen en motorvoertuigen.
In *Tabel 8* en *9* zijn weer de aantallen met de hoogste en de laagste kwaliteitsklasse gegeven. De 44 records die bij de LMR als fiets zijn gecodeerd, maar door de politie als bromfiets, zouden snorfietsen kunnen zijn. *Tabel 9* is op het oog niet van random te onderscheiden, een analyse als in de vorige paragraaf geeft aan dat er circa 140 (5,7%) records bij zijn met hetzelfde patroon als in *Tabel 8*.

VVM	Voet	Fiets	Brom	Motor	Auto	Vr/B	Overig	Onbek.	Totaal
Voet	67	8						5	80
Fiets	11	637	1		3			7	659
Brom	3	44	25						72
Motor		1	1					1	3
Auto		4			1	1			6
Vracht		1			1				2
Bus						2			2
Rail							2		2
Overig		3					1		4
Totaal	81	698	27	0	5	3	3	13	830

Tabel 8. *Vervoerswijze in de VOR (verticaal) tegen die in de LMR, bij ongevallen met overige voertuigen in 1992 en 1993, KOPKWAL 1.*

VVM	Voet	Fiets	Brom	Motor	Auto	Vr/B	Overig	Onbek.	Totaal
Voet	15	148	1		1		1	4	170
Fiets	13	673	5		2	2	8	14	717
Brom	14	431	8	1		1	5	4	464
Motor	1	124	1					1	127
Auto	26	905	9		5	2	12	21	980
Vracht		23							23
Bus		10							10
Rail		2						1	3
Overig	1	6						2	9
Totaal	70	2322	24	1	8	5	26	47	2503

Tabel 9. *Vervoerswijze in de VOR (verticaal) tegen die in de LMR, bij ongevallen met overige voertuigen in 1992 en 1993, KOPKWAL 6.*

De tussenliggende KOPKWAL klassen vertonen eenzelfde overgang naar een random patroon als bij de groep motorvoertuigen. Ook hier kent KOPKWAL 1-3 een verwaarloosbaar aandeel random, en leveren KOPKWAL 4-6 nog 85, 5 en 140 terecht gekoppelde records.

3.6.3. *Overige E-codes*

In de eerste plaats vallen hieronder de groep 'Spoorwegongevallen' (E-code 801, 805-807). Daarvan zijn 10 records - van het totale aantal van 35 in de jaren 1992 en 1993 in de LMR - gekoppeld met KOPKWAL 1, waarvan 7 overeenstemmend (1 voetganger, 5 fietsers en 1 brommer), 2 waren niet gespecificeerd en slechts 1 was fout (waarschijnlijk een codeerfout). In de overige kwaliteitsklassen waren 7 records, waarvan 2 goed, 1 volgens de oude codering, 1 niet gespecificeerd en 3 fout.

Ten tweede is de groep 'Niet-verkeersongevallen met een motorvoertuig' (E820-825, dit zijn volgens de E-code ongevallen *buiten de openbare weg*) apart beschouwd. Deze groep heeft bij de koppeling een afstandbijdrage van 50 gekregen, zodat pas vanaf KOPKWAL 3 records voorkomen. In deze groep hebben van de 106 records 78 overeenstemmende vervoerswijzen, 2 de oude codering, zijn 9 niet gespecificeerd en 17 fout. Het beeld in de overige klassen lijkt op dat bij de ongevallen met overige voertuigen: weinig records in de tussenliggende klassen en een nagenoeg perfecte random tabel (met 123 records) bij KOPKWAL 6.

Ten slotte is een tweetal groepen E-codes toegevoegd die geen directe relatie met verkeer hebben, en dus geen codes voor de vervoerswijze van het slachtoffer (de 'Niet-verkeers-groepen'). De footprint-methode kan hier dus niet toegepast worden (zie ook § 3.4). Het gaat om de 'Zelfmoord-(pogingen)' (E958 en E988) en 'Niet gespecificeerde ongevallen' (E928.9), waarvan de laatste verreweg het grootst is.

Hier (*Tabel 10*) kan, voor de verschillende koppelklassen, alleen de verdeling van de vervoerswijze volgens de VOR gegeven worden, want de vervoerswijze wordt bij deze groepen niet gecodeerd door de LMR. Ook hier werd een bedrag bij de afstand geteld, respectievelijk 90 en 55. Het is

zinvol het totaal te vergelijken met dat van alle goed gekoppelden bij de verkeersongevallen uit de standaard E-codegroep (volgens de LMR). We zien dat de verdelingen redelijk op elkaar lijken, met een wat groter aandeel fietsers en brommers in de niet-verkeers-groep, en een wat kleiner aandeel motoren en auto's.

Door de hogere eisen die bij deze groepen aan de koppeling werden gesteld (door de bijtelling van een afstand) en gezien de uniformiteit van de verdelingen over de vervoerswijzen is het denkbaar dat een minstens even groot aandeel van de mindere kwaliteitsklassen, zoals K-5 en K-6, terecht gekoppeld is als bij de groepen die als verkeersongeval geregistreerd werden. Omdat het precieze aandeel niet op vergelijkbare wijze vastgesteld kan worden is er van afgezien een schatting van de omvang te maken.

KOPKWAL / verv. wijze	K - 3	K - 4	K - 5	K - 6	Totaal K 3 - 6	Standaard-groep LMR
Voetganger	107	28	11	103	249	2111
Fiets	289	72	38	366	765	5121
Brom/snorfiets	172	45	19	322	558	3675
Motor/Scooter	57	16	7	119	199	1647
Auto	404	102	55	666	1227	8801
Vrachtauto	9	4	2	13	28	127
Bus	2		1	13	16	32
Railvoertuigen			1	2	3	8
Overige vvm.	1	1		10	12	76
Totaal	1041	268	134	1614	3057	21598

Tabel 10. *Niet-verkeers-groepen LMR naar KOPKWAL 3-6 en totaal 3 t/m 6, en totaal goed gekoppeld uit de Standaardgroep LMR, naar vervoerswijze volgens de VOR, 1992&1993.*

3.6.4. *Resultaat*

Samenvattend vinden we de in *Tabel 11* opgenomen aantallen terecht gekoppelde records binnen de verschillende E-code klassen, en een totaal van 22078 terecht gekoppelde records binnen de gekoppelde records van de twee bestanden, over 1992 en 1993 tezamen.

KOPKWAL >	K - 1	K - 2	K - 3	K - 4	K - 5	K - 6	Totaal
Motorvoertuig	11304	3453	2907	373	62	919	19018
Ov. voertuigen	830	251	302	85	5	140	1613
Spoorweg	10	1	2	1			14
Niet op. weg			106	18			124
Niet gespec.			1041	268			1309
Totaal	12144	3705	4358	745	67	1059	22078

Tabel 11. *Aantallen terecht gekoppelde records naar E-code klasse en KOPKWAL, 1992&1993.*

Bij de aantallen in deze en volgende tabellen moet bedacht worden dat vele het resultaat zijn van schattingen met onzekerheden die tot tientallen en meer kunnen oplopen. Toch zijn de aantallen altijd gegeven met de resolutie van een geheel aantal slachtoffers.

4. Schatting omvang doelpopulatie

4.1. Inleiding

Zoals in § 3.5 al beschreven is nog een analyse gemaakt volgens de footprint-methode, maar nu voor alle combinaties van afstandklasse en selectiviteitsklasse. De indruk ontstond namelijk dat de groepen met niet te grote afstand maar met kleine selectiviteit (die alle in de groep KOPKWAL 6 zitten) toch in veel gevallen *correct* en wellicht zelfs *terecht* gekoppeld zijn. Daarbij bleek dat voor afstanden tot 65 de footprint niet verschilde van de groepen met grotere selectiviteit.
De conclusie is dat deze groepen beter bij de andere met dezelfde afstand-klasse gevoegd kunnen worden. Voor met kleine selectiviteit gekoppelde records zou het kunnen gebeuren dat twee gelijke records (administratieve meerlingen) verwisseld zijn. Ze zijn dan niet meer terecht gekoppeld, maar hun *groep* is nog wel correct gekoppeld. In de groep met afstand 66-100 bleken de verschillen wat groter te zijn.
Dit heeft ertoe geleid dat voor de definitieve analyse de gekoppelde records met afstand tot 65 (AFKLASC 0-2, zie § 2.1.2) als correct gekoppeld zijn beschouwd. Deze groepen zijn tezamen genomen en dienden als footprint voor de vaststelling van de aandelen correct gekoppelde records in de groepen met grotere afstand.
Vanwege de gevonden overeenkomsten zijn een paar E-code-groepen samengevoegd, maar onder weglating van delen met een zeer lage a priori kans op records uit de doelpopulatie.
Samengevoegd zijn de 'Motorvoertuigongevallen', de 'Spoorweg-ongevallen' en de groep 'Buiten de openbare weg', zonder de E-codes 817 en 818 (dus E801, E805-E807, E810- E816, E819-E825); de groep 'Langzaam-verkeer-ongevallen' is gebruikt zonder E828 en bestaat nu uit E826, E827 en E829. De eerste groep zullen we vanaf nu '*Motorvoertuig*-ongevallen' noemen en de tweede '*Langzaam*-verkeer-ongevallen'.

4.2. Schatting van de doorsnede van LMR- en VOR-bestand

Voor de bepaling van het totaal aantal ziekenhuisgewonden in (een van) beide bestanden, de *vereniging* van beide bestanden, is het nodig te weten hoeveel gevallen in beide voorkomen, de *doorsnede* van de bestanden, omdat deze laatsten in de vereniging enkel geteld moeten worden.

De doorsnede bestaat in ieder geval uit de tot afstand 65 gekoppelde records. Daarin zitten behalve de records uit de eerste drie kolommen (K-1 - K-3) van *Tabel 11* de records met selectiviteit kleiner dan 40 die uit K-6 komen. Daaraan kunnen wij toevoegen de uit bovengenoemde analyse volgende aantallen *correct* gekoppelde records die in de groepen met grotere afstand gekoppelde records te herkennen zijn. (Dit overigens zonder te kunnen vaststellen welke records de terecht gekoppelde zijn.)
Het feit dat er nog flinke aantallen terecht gekoppelde records voorkomen onder de slecht - dus op afstand tussen 100 en 200 - gekoppelde records wijst op de mogelijkheid dat ook onder de *niet* gekoppelde records nog bij elkaar horende paren voorkomen.

Als de maximale grens waarbij gekoppeld wordt, hoger dan 200 gesteld was zouden meer records gekoppeld zijn, waaronder naar alle waarschijnlijkheid ook terecht gekoppelde. Om een indruk te krijgen hoeveel tot de doorsnede behorende records nog in de restbestanden voorkomen is gekeken of de dalende tendens van aandelen correct gekoppelde records bij toenemende afstand kon worden geëxtrapoleerd. Deze gegevens lieten echter niet toe om tot een bijtelling te komen. Bij een toekomstige koppeling kan onderzocht worden om hoeveel het gaat. De resultaten van de analyse staan in *Tabel 12.*

Slachtoffers	A = 0	10 - 40	44 - 65	66 - 100	101 - 130	131 - 200
Motor	11302	3237	3069	785	649	2342
Perc. correct	100%	100%	100%	79%	40%	14%
Langzaam	839	247	310	180	394	2052
Perc. correct	100%	100%	100%	51%	15%	1,6%

Tabel 12. *Aantallen per LMR-groep en afstandklasse en percentages correct gekoppelde records. Jaren 1992 en 1993 gesommeerd.*

Het resultaat is dat uit de groepen met afstand groter dan 65 bij de 'Motorvoertuigongevallen' nog 1205 en bij de 'Langzaam-verkeer-ongevallen' nog 183 records als correct gekoppeld beschouwd worden.
Bij de overige groepen (E817, E818, E828, de 'Niet gespecificeerde' en de 'Zelfmoordongevallen' worden de groepen tot afstand 100 als correct gekoppeld beschouwd. Het resultaat staat in *Tabel 13.*

Slachtoffers	Correct gekoppeld	Totaal gekoppeld
Motorvoertuigongevallen	18813	21402
Langzaam-verkeer-ongevallen	1579	4030
E817, E818 en E828	420	720
Niet gespecificeerde ongevallen	1361	3030
Zelfmoord(poging)	9	28
Totalen	22182	29210

Tabel 13. *Aantallen correct en totaal gekoppelde records naar LMR-groep. Jaren 1992 en 1993 gesommeerd.*

4.3. Schatting van de omvang van de doelpopulatie binnen de restbestanden

4.3.1. *Inleiding*

In het verleden is gebleken dat bij het omschrijven van ongevallen door gewonde verkeersslachtoffers, bij hun contacten met ziekenhuizen, een ruimere groep ongevallen als verkeersongevallen wordt beschouwd dan overeen komt met de (internationale) definitie die in principe door de politie gehanteerd wordt. Het gaat hier onder meer om het onderscheid tussen

openbare en niet-openbare weg. Dat onderscheid wordt binnen de systematiek van de E-code alleen gemaakt bij de ongevallen waarbij een motorvoertuig (inclusief bromfiets) betrokken is.
Bij de groep 'Ongevallen met andere wegvoertuigen' (E-code 826-829), die hoofdzakelijk fietsers betreft, is dat onderscheid er niet en een tot nu toe onbekend deel van die ongevallen valt dus buiten de doelpopulatie. Als het om een aanmerkelijk deel van die ongevallen gaat zou dat mede het grotere aandeel gewonde fietsers in de LMR kunnen verklaren zoals dat uit de gepubliceerde ziekenhuisgegevens naar voren komt. Ook blijken onder sommige E-codes ongevallen gerekend te worden die om andere redenen (grotendeels) niet tot de doelpopulatie behoren. Hieronder vallen E817: 'Verkeersongeval met een motorvoertuig, tijdens het in- en uitstappen, zonder botsing', omdat er geen rijdend voertuig bij betrokken is, en E828: 'Ongeval met een bereden dier', omdat een ruiter bij de VOR tot de voetgangers gerekend wordt en er dus ook in veel gevallen geen rijdend voertuig bij betrokken is.
Van de beschreven E-codes vallen veel records in het restbestand van de LMR. Van de restbestanden uit VOR en LMR moet nu geschat worden welk deel tot de doelpopulatie behoort.
De schatting van de aantallen uit de restbestanden kan minder nauwkeurig gebeuren dan die in de slecht gekoppelde delen van de gekoppelde bestanden, omdat minder informatie voorhanden is, die slechts uit één der bestanden komt. Het principe gaat als volgt. De restbestanden worden onderverdeeld in groepen met verschillende a-priori-kansen op aanwezigheid van records uit de doelpopulatie. Bij de het LMR-restbestand gaat het om de verschillende E-code groepen.
Enerzijds gaat het daarbij om groepen die bij deze koppeling zijn toegevoegd om te onderzoeken of daarin nog een aandeel verkeersslachtoffers voorkomt, namelijk de groep motorvoertuigongevallen buiten de openbare weg, de groep spoorwegongevallen, de zelfmoord(poging)en en verreweg de grootste, de niet gespecificeerde ongevallen.
De eerste twee groepen zijn bij de motorvoertuigongevallen gevoegd omdat ze aan dezelfde analyse onderworpen konden worden doordat de vervoerswijze op dezelfde manier gecodeerd wordt.
Bij de laatste twee groepen bleken goed koppelbare verkeersongevallen voor te komen, maar het betrof slechts een relatief klein deel van die groepen (zie ook deel A, *Tabel 4*). Anderzijds zijn bij delen van de standaardgroep van verkeersongevallen vraagtekens gezet betreffende hun voldoen aan de definitie van verkeersongeval. Dit zijn de codes E817, E818 en E828 die volgens hun definitie niet gepaard gaan met een rijdend vervoermiddel. Deze hebben bij de koppeling dan ook een extra afstand toegekend gekregen van 90. Uit de verdeling van de toch gekoppelde records over de afstandklassen viel op te maken dat ook zonder deze extra afstand niet veel meer van deze records gekoppeld zouden zijn.
Bij deze groepen met lage a-priori-kansen is aangenomen dat de goed (met KOPKWAL 1-4) gekoppelde records terecht gekoppeld zijn, en dat de andere niet tot de doelpopulatie behoren.
Bij de VOR gaat het hier om de groep waar de politie heeft aangegeven dat er geen ziekenhuisopname was, en waarvan slechts ruim 3% goed gekoppeld kon worden.

Anders ligt het bij de groepen met hoge a-priori-kans op behoren tot de doelpopulatie. De grootste hierbij is het restbestand bij de LMR dat voor

meer dan de helft bestaat uit fietsongevallen. Bij deze groep is op basis van de vergelijking van relevante verdelingen tussen restgroep en gekoppelde groep en van kennis uit andere bronnen een schatting gemaakt van het aandeel dat tot de doelpopulatie behoort.

4.3.2. *Het restbestand VOR*

Op basis van de berekende doorsnede van beide bestanden is de verdeling berekend binnen het VOR-bestand en weergegeven in *Tabel 14*. In de rest-bestanden komen 5.354 records voor waarbij gecodeerd is dat opname in een ziekenhuis heeft plaatsgevonden (van de andere wordt aangenomen dat inderdaad niet van opname sprake was). Het niet terugvinden van deze records in de LMR kan twee hoofdoorzaken hebben.

Slachtoffers	Doorsnede	Rest	Totaal
Doden (2-30 dagen na ongeval)	399	70	469
Opgenomen	17866	5354	23220
Zelfde dag overleden	200	421	621
Naar ziekenhuis, opname onbekend	1210	3522	4732
Naar ziekenhuis, niet opgenomen	1423	31281	32704
Niet naar ziekenhuis & alles onbekend	1070	34088	35158
Ter plaatse overleden	14	1433	1447
Totalen	22182	76169	98351

Tabel 14. *Verdeling van het koppelresultaat naar ernstgroep VOR, 1992 & 1993.*

Ten eerste kan het zijn dat de codering juist is maar dat het slachtoffer niet in het LMR-bestand terecht is gekomen. Dit kan volgens opgave van de SIG (de schatting van de percentages is in overleg met de SIG gebeurd; ze slaan op het VOR-bestand van opgenomenen over 1992 en 1993) gebeurd zijn doordat:

- de opname in een buitenlands ziekenhuis was (0,2%, 50);
- de opname door het ziekenhuis niet bij de LMR geregistreerd is (1,5%, 360);
- de patiënt nog niet ontslagen was (enkele gevallen per jaar);
- er geen of verkeerde E-code gebruikt is (2%, 480).

Ten tweede kan de codering onjuist zijn: de politiefunctionaris nam aan dat het slachtoffer was opgenomen maar deze opname heeft niet plaats-gevonden. Dat het hierbij om een aanzienlijk aantal slachtoffers kan gaan werd al eerder aangenomen, omdat de politie niet bij de feitelijke opname aanwezig is.
Ook ander onderzoek wijst in dezelfde richting. Bij een in 1989 uitgevoerde proef, waarbij aan het al bestaande Privé-ongevallenregistratiesysteem (PORS) de verkeersongevallen werden toegevoegd, is ook een handmatige koppeling van de verkeersslachtoffers met de VOR uitgevoerd (Blokpoel, 1990). Daarbij bleek dat het gegeven 'opgenomen in een ziekenhuis' in een aanmerkelijk deel van de gekoppelde gevallen niet met de gegevens uit het

ziekenhuis overeen kwam. In 13% van de gevallen waarbij de politie aangaf dat opname had plaatsgevonden bleek dit niet het geval te zijn, terwijl in 7% van de gevallen dat volgens de politie geen sprake was van opname dit niet klopte met de PORS-gegevens.
We zien eenzelfde asymmetrie als bij dit onderzoek, alleen minder extreem (zie bijvoorbeeld *Tabel 1*, waaruit blijkt dat in 1992 26,3% van de volgens de politie opgenomen slachtoffers niet of slecht gekoppeld kon worden, terwijl van de categorie 'Niet opgenomen' nog 3,3% wel goed koppelbaar bleek).

Er resulteert dus een groep van 4.464 (5354-50-360-480), wat neerkomt op ruim 18% van alle (volgens de VOR) opgenomen slachtoffers. Op basis van de te onzer beschikking staande gegevens wordt verder aangenomen dat deze groep niet tot de doelpopulatie behoort. Het belangrijkste argument daarbij is de overweging dat de politie, na waargenomen te hebben dat een slachtoffer gewond is en naar een bepaald ziekenhuis vervoerd is, niet meer van de gebeurtenissen in het ziekenhuis op de hoogte hoeft te zijn. Wordt iemand na uitvoerige behandeling op de eerste hulp opgenomen of kan hij of zij toch naar huis? Het onderscheid tussen beide behandelingen is voor de afhandeling van de politie van minder belang dan andere - meer verkeers-technische - gegevens.

4.3.3. *Het restbestand LMR*

Van de restbestanden zijn de bovenste twee uit groepen met hoge a-priori-kans op behoren tot de doelpopulatie, van de overige groepen wordt aangenomen dat de restbestanden geen verkeersslachtoffers zijn.

Slachtoffers bij:	Doorsnede	Rest	Totaal
Motorvoertuigongevallen	18813	8129	26942
Langzaam verkeer ongevallen	1579	8274	9853
E817, E818, E828	420	2286	2706
Niet gespecificeerde ongevallen	1361	11763	13124
Zelfmoord(pogingen)	9	329	338
Totalen	22182	30781	52963

Tabel 15. *Verdeling van het koppelresultaat naar ernstgroep LMR, 1992 & 1993.*

Een wat groter deel (8.274) wordt gevormd door de groep ongevallen waarbij geen motorvoertuig (inclusief bromfiets) betrokken was (E826, E827, E829). Het slachtoffer was in verreweg de meeste gevallen fietser. Ook hier zijn weer twee hoofdoorzaken te geven waarom het slachtoffer niet in het andere bestand (VOR) te vinden is, namelijk het behoort niet tot de doelpopulatie, of het ontbreekt in het andere bestand.

Ten eerste kan de codering onjuist zijn: het was geen verkeersongeval volgens de internationale definitie. Bij de groep 'Langzaam verkeer' wordt niet onderscheiden naar wel of niet openbare weg, maar ongevallen buiten

de openbare weg voldoen niet aan de internationale definitie en vallen dus buiten de doelpopulatie. Uit drie bronnen kan een schatting gemaakt worden van het aandeel van deze groep dat buiten de openbare weg plaatsvond.
De ongevallen met motorvoertuigen worden wel onderscheiden naar al of niet openbare weg. Hier ligt de verhouding zo dat 2,9% als buiten de openbare weg geregistreerd wordt. Uit het feit dat van de records die als op de openbare weg gebeurd gecodeerd zijn een hoog percentage goed gekoppeld is (69,4%) terwijl van de groep buiten de openbare weg slechts 15,2% goed koppelbaar was kan opgemaakt worden dat het gegeven 'al of niet openbare weg' redelijk goed gecodeerd wordt. Daardoor is het eerder genoemde percentage buiten de openbare weg als betrouwbaar te beschouwen.
Het lijkt aannemelijk een iets hoger percentage te hanteren voor de groep E826-E829, omdat langzaam fietsverkeer zich wat vaker buiten de openbare weg kan afspelen dan gemotoriseerd verkeer.
Bij het eerder genoemde proefonderzoek PORS kwam ook een zeer groot aandeel fietsslachtoffers voor. Bij deze groep is nagevraagd of het ongeval al of niet op de openbare weg had plaatsgevonden. Dit was in circa 4% van de gevallen buiten de openbare weg.
Ook in het onderzoek Ongevallen in Nederland (OIN) (SWOV, 1995), een nu tweemaal gehouden enquête onder een steekproef uit alle Nederlanders is naar dit gegeven gevraagd. Van de relevante groep ongevallen (met ziekenhuisopname) bleek 4 à 5% buiten de openbare weg te vallen.
Daarmee is het aannemelijk dat circa 5% van de totale groep E826, E827 en E829 (493 records) als niet tot de doelpopulatie behorend kunnen worden opgevat.
Ook kunnen records voorkomen waarbij ten onrechte een code uit de standaardreeks is gecodeerd. Het was bijvoorbeeld geen verkeersongeval maar een privé-ongeval. Volgens opgave van de SIG zou het om circa 0,2% (76) van het totaal kunnen gaan, maar het omgekeerde komt waarschijnlijk even vaak voor.

Ten tweede kan het niet aanwezig zijn van de tegenhanger van een LMR-record in het VOR-bestand zijn oorzaak vinden in het feit dat het in het VOR-bestand ontbreekt. Hiervoor bestaan volgens opgave van AVV/BG weer verschillende mogelijk oorzaken (geldig voor ziekenhuisgewonden):
- het ongeval is in het buitenland gebeurd (0,2%, 50);
- het ongeval miste locatiegegevens (0,2%, 50);
- het ongevalsformulier werd te laat ingezonden (naijlers) (circa 2%, 500);
- het ongeval is niet door de politie aan de VOR gemeld (respectievelijk was niet bij hen bekend).

In het eerste geval behoort het niet tot de doelpopulatie, in de andere gevallen wel, als aangenomen mag worden dat de LMR-codeur terecht een code uit de standaardgroep heeft gekozen. Op basis van de eerder genoemde onderzoeken PORS en OIN, en een latere uitbreiding van PORS met verkeersongevallen (VIPORS) is het aannemelijk dat het overgrote deel van deze ongevallen aan de definitie voldeed.
Een belangrijk deel vormen ongevallen waarbij iemand met zijn fiets ten val komt of tegen iets aanrijdt. Ook botsingen tussen fietsers onderling en tussen fietsers en voetgangers komen voor. Het is begrijpelijk dat de politie van een veel kleiner aandeel van deze ongevallen in kennis gesteld wordt dan van ongevallen met motorvoertuigen. Het aandeel niet urgente opnames is in deze groep groter, wat wijst op minder ernstige verwondingen. Er is

geen verplichte verzekering in het spel, de materiële schade is meestal veel geringer en ze zullen vaker gebeuren in een woonomgeving, die hulp van de politie bij het medisch begeleiden tot aan het ziekenhuis minder nodig maakt.

4.4. Resultaat

Uit het voorgaande zijn voor beide bestanden de meest aannemelijke schattingen bepaald van de omvang van tot de doelpopulatie behorende delen. Alle records zijn vervolgens verdeeld naar voorkomend in de doorsnede, wel behorend tot de doelpopulatie (maar niet in de doorsnede) en niet behorend tot de doelpopulatie.
De resultaten voor de LMR zijn te vinden in *Tabel 16*. De cel 'Wel in VOR - Niet in LMR' is in § 4.3.2 behandeld. In *Tabel 17* is *Tabel 4* ingevuld.

Slachtoffers bij:	Doorsnede	Wel doelpopulatie	Geen doelpopulatie	Totaal
Motorvoertuigongevallen	18813	8037	92	26942
Langzaam verkeer ongevallen	1579	7747	527	9853
E817, E818, E828	420	-	2286	2706
Niet gespecificeerde ongevallen	1361	-	11763	13124
Zelfmoord(pogingen)	9	-	329	338
Totalen	22182	15784	14997	52963

Tabel 16. *Verdeling van het koppelresultaat naar ernstgroep LMR, 1992 & 1993, naar doorsnede en doelpopulatie.*

Slachtoffers	Wel in LMR	Niet in LMR	Geen doelpopulatie
Wel in VOR	22182	890	75279
Niet in VOR	15784	In geen van beide	
Geen doelpopulatie	14997		

Tabel 17. *Slachtoffers in de VOR (horizontaal) en de LMR (verticaal), 1992 & 1993 naar overlap en doelpopulatie.*

4.5. Slachtoffers die in geen van beide bestanden voorkomen

Omdat in beide bestanden delen bestaan die wel tot de doelpopulatie behoren, maar niet in het andere bestand voorkomen, bestaat de mogelijkheid dat er slachtoffers bestaan die in beide bestanden ontbreken. Als aangenomen wordt dat de processen, die in elk der bestanden leiden tot het in de registratie ontbreken van ziekenhuisgewonden, onafhankelijk van elkaar werken volgt het aantal door een proportioneel deel te plaatsen in de cel 'In geen van beide' van *Tabel 17*.
Er is op basis van de beschikbare kennis geen reden om aan te nemen dat de verschillende eerder genoemde redenen, waardoor slachtoffers niet in een der bestanden geregistreerd worden, sterke afhankelijkheden zullen

vertonen. Daaruit volgt voor deze cel een aantal van 633 slachtoffers (15784 * 890 / 22182). Het resultaat is te vinden in *Tabel 18*, die de invulling is van *Tabel 3*.

Doelpopulatie	Wel in LMR	Niet in LMR	Totaal
Wel in VOR	22182	890	23072
Niet in VOR	15784	633	16417
Totaal	37966	1523	39489

Tabel 18. *Verdeling doelpopulatie naar voorkomen in VOR en/of LMR, 1992 & 1993.*

4.6. Conclusie

Het bovenstaande leidt tot een totaal generaal van 39.489 verkeers-slachtoffers die in een ziekenhuis zijn opgenomen in de jaren 1992 en 1993. Dit is circa 2% meer dan het aantal uit de standaardgroep van de LMR: 38.646.
Alles bij elkaar genomen is het resultaat dat het totaal geschatte aantal verkeersslachtoffers (volgens de internationale definitie) dat in een ziekenhuis is opgenomen - de doelpopulatie - niet veel verschilt van het door de LMR jaarlijks gevonden aantal uit de standaardreeks. Dit is het resultaat van enerzijds het verwijderen uit de standaardreeks van groepen die bij nader inzien niet tot de doelpopulatie behoren en het toevoegen van delen die van buiten de standaardreeks komen.
Eenzelfde beeld van toevoegen en verwijderen bestaat bij de VOR als het gaat om de slachtoffers van ongevallen waarbij motorvoertuigen betrokken zijn. Het bestaan van een grote groep slachtoffers van - meest eenzijdige - fietsongevallen die slechts in geringe mate door de politie worden geregistreerd is bevestigd.

5. Het berekenen van ophoogfactoren

5.1. Inleiding

In principe kan vanuit ieder der twee bestanden, LMR en VOR, opgehoogd worden naar het geraamde totaal van de doelpopulatie, als (nog) geen koppelgegevens beschikbaar zijn. Omdat het principe gelijk is maar de LMR relatief weinig slachtoffers mist, de VOR de officiële registratie van verkeersongevallen is en meestal het eerst beschikbaar is, wordt hier verder gesproken over het ophogen van VOR-cijfers naar geraamde totalen.
Na het beschikbaar komen van de LMR-cijfers kunnen die ook - in geringe mate - opgehoogd worden (met circa 2%) en moet de (LMR-)verdeling over de vervoerswijzen op de hieronder beschreven wijze gecorrigeerd worden zodat hij overeenstemt met de (juist geachte) VOR-indeling. Uit deze koppeling is immers een goed beeld verkregen over de verschillen tussen die twee coderingen.
Een belangrijk punt is of de ophoging plaatsvindt vanuit de aantallen binnen de formele doelpopulatie van het bestand, bij de VOR dus de opgenomenen, of dat vanuit andere groepen - zoals bijvoorbeeld - *wel naar ziekenhuis maar opname onbekend* - een eigen deel wordt opgehoogd. Dit zou in ieder geval moeten als de verhoudingen wel/niet doelpopulatie in de groepen opgenomen en opname onbekend in de tijd zouden veranderen.
Een vergelijking tussen 1992 en 1993 wijst niet op zo'n verandering.
Het verdient dan wel aanbeveling de onveranderlijkheid van deze verhoudingen te onderzoeken door de koppeling om de paar jaar te herhalen.
Het eenvoudigst zou zijn om de ophoging te baseren op de in de bestanden aanwezige delen die tot de formele doelpopulatie behoren. Voor de ophoging naar onderverdelingen gelden nieuwe eisen: zo moeten de verhoudingen naar formele doelpopulatie versus andere groepen niet teveel verschillen. Bij de bovengenoemde verdeling naar *opgenomen*, *opname onbekend* en *niet opgenomen* ontstaat dan bijvoorbeeld een probleem doordat in een aantal provincies *opname onbekend* praktisch niet voorkomt. Dit leidt tot grotere onzekerheid in de opgehoogde aantallen.

De ophoogfactoren geven de feitelijke verhoudingen weer tussen de aantallen in het VOR-bestand van ziekenhuisgewonden en de geraamde totalen, over 1992 en 1993. Toepassing van deze factoren op VOR-gegevens uit andere jaren dan 1992/93 geeft alleen cijfers zonder systematische fouten onder de aanname dat de registratieprocessen bij de politie en/of AVV/BG niet veranderen. Bij veranderingen in deze processen zal de koppeling zeker herhaald moeten worden.

In het vervolg van dit hoofdstuk zullen de ophoogfactoren berekend worden, verdeeld naar vervoerswijze en provincie.

5.2. Wijze van verkeersdeelname

Bij deze zeer belangrijke variabele treden meteen een aantal problemen naar voren, door de grote verschillen tussen de wijze waarop hij verdeeld is bij VOR en LMR. Zoals bekend komen in de LMR veel meer fietsslachtoffers

voor dan in de VOR, maar het zou onjuist zijn bij de ophoging naar vervoerswijze de LMR-verdeling aan te houden, omdat we - mede door deze koppeling - weten dat bij de codering van deze variabele in de ziekenhuizen fouten worden gemaakt, terwijl van een niet onbelangrijk deel de vervoerswijze als onbekend wordt gecodeerd.
Met de zeer goed gekoppelde records (KOPKWAL 1) beschikken we echter over veel informatie over de werkelijke verdeling over de vervoerswijzen, als tenminste wordt aangenomen dat de politie dit gegeven foutloos codeert. Dit betekent dat voor de vier cellen in *Tabel 18*, die tezamen de doelpopulatie omvatten, de verdeling over de vervoerswijzen aangegeven moet worden. Het grootste deel wordt gevormd door de doorsnede, de groep die in beide bestanden voorkomt. Daarvan is zowel de verdeling volgens de LMR als die volgens de VOR bekend.
Binnen die doorsnede zijn weer twee sterk verschillende delen te onderscheiden, de slachtoffers van ongevallen met motorvoertuigen en die met alleen overige voertuigen. In *Tabel 6* en *8* zijn de relaties tussen de twee codeerwijzen goed te zien. Het eenvoudigst zou zijn dezelfde relaties te gebruiken binnen de groep 'wel in LMR - niet in VOR' ('Wel doelp.'), maar dan moet de verdeling binnen die groep volgens de LMR niet te veel verschillen van die in de doorsnede.
In *Tabel 19* zijn de verdelingen gegeven voor de motorvoertuigongevallen bij de LMR, en in *Tabel 20* die voor de groep ongevallen met andere wegvoertuigen. De verdelingen zijn in procenten gegeven om te benadrukken dat het uit de totalen *berekende* aandelen zijn, waarbij de verdeling van de Doorsnede genomen is uit de footprint-tabel en die van de groep 'Wel doelp.' uit de betreffende restgroep. De totalen komen uit *Tabel 16*.

Percentages	Voet	Fiets	Brom	Motor	Auto	Vr/B	Overig	Onb.	N tot.
Doorsnede	12,3	15,9	14,8	7,2	38,3	1,1	0,9	9,6	18813
Wel doelp.	13,4	12,3	16,3	11,4	30,8	2,6	1,5	11,8	8037

Tabel 19. *Verdeling naar LMR-vervoerswijze bij de doorsnede en de groep wel doelpopulatie/geen doorsnede in de LMR, 1992&1993, motorvoertuigongevallen exclusief E817 en E818.*

Percentages	Voet	Fiets	Brom	Motor	Auto	Vr/B	Overig	Onb.	N tot.
Doorsnede	9,3	85,0	3,0	0,1	0,7	0,3	0,3	1,5	1579
Wel doelp.	2,9	94,0	0,8	0,1	0,4	0,1	1,0	1,1	7747

Tabel 20. *Verdeling naar LMR-vervoerswijze bij de doorsnede en de groep wel doelpopulatie/geen doorsnede in de LMR, 1992&1993, ongevallen met andere wegvoertuigen exclusief E828.*

Hoewel de verschillen niet zeer groot zijn, moet er toch rekening mee gehouden worden. Hiertoe moet voor de twee groepen een transformatietabel opgesteld worden die de LMR-vervoerswijze omzet naar de 'werkelijke', op de VOR-gegevens gebaseerde vervoerswijze. Dit is gedaan met behulp van de tabellen naar vervoerswijze zoals *Tabel 6*, maar dan voor de afstandsklassen 0 t/m 2.

Daartoe zijn eerst van deze tabellen de percentages verticaal berekend. Deze percentages in de kolommen bij de LMR-vervoerswijzen, bijvoorbeeld 'Fiets', geven aan in welk deel daarvan het in werkelijkheid ging om voetganger, fietser enzovoort.
Op dezelfde wijze worden nu de aantallen in de restbestanden per LMR-vervoerswijze over de VOR-vervoerswijzen verdeeld. Daarbij zijn vanzelf ook de als 'Onbekend' gecodeerde gevallen ingedeeld. Dit komt neer op matrixvermenigvuldiging van de aantallen behorend bij de onderste regels van *Tabel 19* en *20* als (kolom)vector met de matrix van de kolompercentages.
Op deze wijze zijn de volgende aantallen voor de groepen 'Wel doelp.' verkregen. Voor de doorsnede volgen die aantallen direct uit dezelfde tabellen. Hier zijn wel aantallen gegeven omdat die voor de berekeningvan ophoogfactoren nodig zijn.

Slachtoffers		Voet	Fiets	Brom	Motor	Auto	Vr/B	Overig	Totaal
Doorsnede	N	1846	3638	3167	1492	8476	126	68	18813
	%	9,8	19,3	16,8	7,9	45,1	0,7	0,4	100
Wel doelp.	N	885	1348	1535	960	3185	89	35	8037
	%	11,0	16,8	19,1	11,9	39,6	1,1	0,4	100

Tabel 21. *Verdeling naar VOR-vervoerswijze bij de doorsnede en de rest (wel doelpopulatie/geen doorsnede LMR) in 1992&1993, groep motorvoertuigongevallen exclusief E817 en E818.*

Slachtoffers		Voet	Fiets	Brom	Motor	Auto	Vr/B	Overig	Totaal
Doorsnede	N	153	1248	138	9	15	5	11	1579
	%	9,7	79,0	8,7	0,6	0,9	0,3	0,7	100
Wel doelp.	N	309	6671	552	39	67	13	96	7747
	%	4,0	86,1	7,1	0,5	0,9	0,2	1,2	100

Tabel 22. *Verdeling naar VOR-vervoerswijze bij de doorsnede en de rest (wel doelpopulatie/geen doorsnede LMR) in 1992&1993, groep ongevallen met andere wegvoertuigen exclusief E828.*

Deze verdelingen zijn zoals gezegd anders dan die in *Tabel 19* en *20* door de systematisch andere codeerwijze van het vervoermiddel bij de LMR.

Voor de niet-verkeersgroepen zijn de aantallen direct uit het gekoppelde bestand verkregen:

Slachtoffers	Voet	Fiets	Brom	Motor	Auto	Vr/B	Overig	Totaal
E817-828	24	23	179	110	70	10	4	420
Niet gesp.	141	370	230	76	525	16	3	1361
Zelfmoord	2	1			6			9

Tabel 23. *Verdeling naar VOR-vervoerswijze bij de doorsnede in 1992 & 1993, niet-verkeersgroepen.*

Als laatste moeten de verdelingen verkregen worden van de twee groepen uit *Tabel 18* die niet in de LMR voorkomen, de 890 slachtoffers die wel tot de doelpopulatie binnen de VOR behoren, maar die in de LMR ontbreken, en de 633 slachtoffers die in beide bestanden (geacht worden te) ontbreken. Van de eerste is het aannemelijk dat de verdeling gelijk is aan die van de doorsnede omdat geen reden bestaat waarom deze groep in samenstelling zou afwijken. Over de groep die in beide bestanden ontbreekt is het minst bekend, maar we nemen aan dat hij lijkt op de groep die wel in de LMR-doelpopulatie valt, maar niet in de VOR. Het resultaat is weergegeven in *Tabel 24.*

Slachtoffers	Voet	Fiets	Brom	Motor	Auto	Vr/B	Overig	Totaal
Wel doelp.	87	212	149	68	365	6	3	890
Geen v. beide	48	322	84	40	130	4	5	633

Tabel 24. *Verdeling naar VOR-vervoerswijze bij groepen 'Wel doel-populatie/geen doorsnede VOR' in 1992 & 1993, en 'In geen van beide'.*

Alles bij elkaar genomen geeft dit de volgende tabel:

Slachtoffers	Voet	Fiets	Brom	Motor	Auto	Vr/B	Overig	Totaal
Doorsnede	2166	5280	3714	1687	9092	157	86	22182
Wel L niet V	1194	8019	2087	999	3252	102	131	15784
Wel V niet L	87	212	149	68	365	6	3	890
Geen v. beide	48	322	84	40	130	4	5	633
Totaal	3495	13833	6034	2794	12839	269	225	39489

Tabel 25. *Verdeling naar VOR-vervoerswijze van de totale doelpopulatie naar aanwezig zijn in VOR en LMR, 1992 & 1993.*

Het is zinvol de verdeling van de totale doelpopulatie onder die van de groep opgenomen van de VOR te zetten. De quotiënten geven dan de ophoogfactoren.

Slachtoffers	Voet	Fiets	Brom	Motor	Auto	Vr/B	Overig	Totaal
VOR opgenomen	2322	5558	4061	1769	10110	189	121	24130
Totaal doelpopulatie	3495	13833	6034	2794	12839	269	225	39489
Ophoogfactoren	1,51	2,49	1,49	1,58	1,27	1,42	1,86	1,637

Tabel 26. *Verdeling naar VOR-vervoerswijze van de totale doelpopulatie en die van de groep opgenomen in de VOR, met ophoogfactoren.*

5.3. Provincie

De analyse naar provincie stuit meteen op het probleem dat de variabele OPGEN, waarin de politie aangeeft of opname ja of nee heeft plaatsgevonden, dan wel dat opname onbekend is, sterk verschillende aandelen onbekend te zien geeft bij de verschillende provincies (zie *Tabel 27*). Opvallend in *Tabel 27* is het praktisch afwezig zijn van 'Opname onbekend' in de provincies Groningen, Drenthe en Zeeland. De ophoging voor heel Nederland is gebaseerd op het deel met opname *ja*, met de aanname dat het deel met opname *onbekend* een klein, constant, deel van het geheel uitmaakt. Omdat dit per provincie in het licht van onderstaande cijfers twijfelachtig is kunnen provinciale ophoogfactoren niet op dezelfde wijze berekend worden als voor heel Nederland is gebeurd.
Er bestaat nog een tweede reden waarom zo'n analyse onmogelijk is, namelijk het (provincie)grens-overschrijdende gewondentransport. Voor heel Nederland speelt dit probleem een zeer geringe rol, maar het is bekend dat voor veel plekken in Nederland het dichtstbijzijnde in aanmerking komende ziekenhuis in een andere provincie ligt. Omdat bij de LMR de provincie van het ongeval niet geregistreerd wordt is de voor Nederland gebruikte methode onuitvoerbaar.

	Opgenomen in een ziekenhuis?			
Slachtoffers VOR	Ja	Onbekend	Nee	Totaal
Groningen	960	3	2324	3287
Friesland	861	408	1326	2595
Drenthe	947	33	1640	2620
Overijssel	1820	115	4621	6556
Gelderland	3210	598	8053	11861
Utrecht	1617	371	5180	7168
Noord-Holland	3839	1230	13208	18277
Zuid-Holland	4014	2002	14584	20600
Zeeland	734	11	1429	2174
Noord-Brabant	3946	761	10363	15070
Limburg	1775	475	4431	6681
Flevoland	407	188	867	1462
Totaal	24130	6195	68026	98351

Tabel 27. *Verdeling van de variabele OPGEN naar provincie, VOR 1992 & 1993.*

Het is echter toch mogelijk om van de koppelresultaten gebruik te maken en provinciale ophoogfactoren te schatten als aangenomen wordt dat in iedere provincie dezelfde verhouding bestaat tussen de aantallen *goed gekoppelde* VOR-records en het werkelijke aantal opgenomenen. Dan kan de landelijke waarde van die factor gebruikt worden per provincie.
Voor het aannemen van die vaste verhouding pleit dat we gezien hebben dat landelijk het aantal slachtoffers van ongevallen waarbij een motorvoertuig

(inclusief bromfiets) betrokken is slechts in geringe mate opgehoogd hoefde te worden. Dit maakt aannemelijk dat de politie overal in Nederland dit soort ongevallen, met een zodanige ernst dat ziekenhuisopname volgt, grotendeels registreert. De grootste ophoging vindt plaats bij de groep langzaam verkeer-ongevallen.

Onze aanname houdt nu in dat de verhouding van het aantal slachtoffers van deze laatste groep ongevallen tot dat van de motorvoertuigongevallen in alle provincies (nagenoeg) gelijk is. Het zou vanzelfsprekend goed zijn deze aanname te toetsen. Hiertoe zou eerst onderzocht moeten worden of bij de goed gekoppelde records de woonprovincie van het slachtoffer overeenstemt met die van de plaats van het ongeval. Indien dat in voldoende mate het geval is kan de woonprovincie de plaats innemen van plaats van het ongeval, zodat de landelijke analysemethode toch toegepast kan worden. Deze analyse kan alleen bij de SIG uitgevoerd worden omdat de benodigde variabele om redenen van privacybescherming niet in het geanonimiseerde LMR-bestand is opgenomen.

In *Tabel 28* zijn de aantallen doelpopulatie per provincie berekend door de aantallen goed gekoppeld met de landelijke factor 39.489/21.759 te vermenigvuldigen. De ophoogfactoren volgen dan uit de verhouding doelpopulatie / opgenomen.

	Goed gekoppeld	Doelpopulatie	Opgenomen	Ophoogfactor
Groningen	763	1385	960	1,44
Friesland	988	1793	861	2,08
Drenthe	768	1394	947	1,47
Overijssel	1681	3051	1820	1,68
Gelderland	2903	5268	3210	1,64
Utrecht	1474	2675	1617	1,65
Noord-Holland	3545	6433	3839	1,68
Zuid-Holland	3484	6323	4014	1,58
Zeeland	530	962	734	1,31
Noord-Brabant	3496	6344	3946	1,61
Limburg	1741	3160	1775	1,78
Flevoland	386	701	407	1,72
Totaal	21759	39489	24130	1,637

Tabel 28. *Geschatte doelpopulatie uit aantallen goed gekoppeld en ophoogfactoren na vergelijking met opgenomenen VOR 1992 & 1993.*

Opvallend is de hoge factor voor Friesland. Deze hangt samen met de sterk afwijkende verdeling naar wel/onbekend/niet opgenomen in die provincie.

5.4. Andere indelingen

De botspartner wordt in de LMR alleen onderscheiden naar motorvoertuig (inclusief bromfiets), overige wegvoertuigen (waaronder fietsen en trams) en obstakels. Het is daarom niet mogelijk de veel gedetailleerder verdeling bij de VOR op te hogen.

Een alternatief zou zijn te pogen de VOR-indeling te vertalen naar de indeling zoals de LMR die hanteert. Daarbij doet zich echter het probleem voor dat de indeling zoals bij de LMR gebruikt wordt, een veel te smalle basis geeft voor de groep *langzaam* (ongevallen waarbij geen motorvoertuig betrokken is), omdat voor de grootste groep, de fietsers, alle botspartners onder slechts één code vallen. De groep motorvoertuigongevallen wordt door de VOR goed gerepresenteerd, zodat de verdeling naar botspartner binnen die groep uit de VOR verkregen kan worden.

De verdeling naar leeftijd hangt zo sterk samen met die naar vervoerswijze dat hij als afzonderlijke tabel weinig inzicht geeft.

6. Bruikbaarheid van de ophoogfactoren voor het beleid

6.1. Inleiding

De verkregen ophoogfactoren geven de verhoudingen weer tussen de aantallen in de VOR die formeel tot de doelpopulatie behoren, slachtoffers die volgens de politie in een ziekenhuis zijn opgenomen, en de op basis van de VOR-LMR-koppeling geschatte totale omvang van de doelpopulatie.
De inverse van de ophoogfactoren kunnen opgevat worden als (geschatte) registratiegraden. Zij gelden voor de jaren 1992 en 1993, waarop de koppeling gebaseerd is.
Een belangrijk resultaat is dat de totale omvang van de doelpopulatie weinig verschilt van die van de standaardgroep uit de LMR. Voor het eerst kan dat totaal nu verdeeld worden naar de vervoerswijze volgens de VOR, die als juist opgevat wordt omdat hij door de politie is vastgesteld. Het bestaande beeld is bevestigd dat de VOR-registratie een redelijke afspiegeling vormt van de ongevallen waarbij gemotoriseerde vervoermiddelen bij betrokken zijn, terwijl van de ongevallen waarbij alleen langzaam verkeer betrokken is een belangrijk deel gemist wordt. Dit resulteert in de belangrijk hogere ophoogfactor voor fietsers dan voor alle overige vervoerswijzen.

6.2. Nauwkeurigheid

Iedere ophoogfactor is het quotiënt van twee aantallen: de teller is een geraamd deel van de doelpopulatie, de noemer het overeenkomstige deel uit het VOR-bestand van opgenomen slachtoffers, voor de periode waarover hij berekend is.
Onder de nauwkeurigheid van een ophoogfactor kunnen een aantal aspecten onderscheiden worden. Ten eerste hangt hij voor de periode waarover hij berekend is af van de mate waarin de teller het werkelijke aantal in de doelpopulatie benadert. De fout bij die raming bestaat uit een toevallig (random) deel en een systematische afwijking. Door de grote aantallen waarmee gewerkt wordt is de toevallige fout zeker klein ten opzichte van de systematische fout. Deze laatste is uit de aard der zaak moeilijk in te schatten omdat een onafhankelijke precieze waarde van (delen van) de doelpopulatie ontbreekt.
Bij de motorvoertuigongevallen zal de systematische fout geringer zijn dan bij de langzaam-verkeer-ongevallen omdat LMR en VOR hier redelijk overeenstemmen. Een op kennis van beide registratieprocessen gebaseerde schatting van de systematische fout in de geraamde aantallen doelpopulatie is circa 10% naar boven en beneden voor de motorvoertuigongevallen en circa 20% voor de groep langzaam. Deze fout geeft geen probleem bij vergelijking van opvolgende jaren omdat hij constant is, mits de LMR-registratie niet wijzigt.
Voor de noemer hangt de nauwkeurigheid van opgehoogde waarden van latere jaren af van de omvang van statistische fluctuaties en de stabiliteit van het registratieproces. Bij ophoging vanuit kleine aantallen gaan de fluctuaties in de aantallen slachtoffers een rol spelen. De onnauwkeurigheid in procenten is dan ongeveer $100/\sqrt{n}$, met n het aantal in de noemer.
Als n = 100 heeft de daarop gebaseerde ophoogfactor dus een standaard meetfout van de orde 10%.

6.3. Verloop in de tijd

Als de over 1992 en 1993 berekende ophoogfactoren worden toegepast op latere jaren gebeurt dat in de veronderstelling dat de registratiegraad niet verandert. Dit is in de afgelopen 20 jaar wel gebeurd. In 1974 was het verschil in aantal tussen de VOR en de LMR veel kleiner dan nu: de VOR bevatte circa 84% van de records van de LMR. Dat percentage is over de jaren heen gedaald tot ruim 60% in 1992. Er zijn aanwijzingen dat de registratiegraad sindsdien niet verder is gedaald.
Een controle op deze veronderstelde constantheid is mogelijk als over hetzelfde jaar de LMR-cijfers beschikbaar zijn. Als de totale opgehoogde omvang verschilt van het met 2% vergrote LMR-aantal geeft dit aanleiding wijzigingen in de registratie te vermoeden, bij VOR en/of LMR.
Het verdient aanbeveling om, naarmate dergelijke verschillen groter zijn en zich vaker voordoen, de koppeling vaker te herhalen.
Dit kan in 1997 actueel worden doordat niet uitgesloten is dat de politie in 1996 anders is omgegaan met de codering van de opname in het ziekenhuis: het VOR-bestand over 1996 bevat relatief meer slachtoffers met 'opgenomen' en minder slachtoffers met 'opname onbekend' dan de bestanden van vorige jaren.

6.4. Beleidsaanbevelingen

Omdat de opgehoogde aantallen slachtoffers een betere benadering zijn van het werkelijke aantal in een ziekenhuis opgenomen slachtoffers, bevelen wij aan in het verkeersveiligheidsbeleid met deze opgehoogde aantallen te werken en ze te beschouwen als de 'officiële' landelijke cijfers. De LMR dateert al van voor 1985 en de taakstellingen voor de jaren 2000 en 2010 kunnen dus, in elk geval voor Nederland als geheel, eveneens worden opgehoogd.
De ophoogfactoren variëren sterk over de provincies. Wij bevelen aan deze uitkomsten aan de registrerende instantie, de politie, voor te leggen om meer inzicht te krijgen in de betekenis van de geconstateerde verschillen. Het lijkt erop dat bepaalde korpsen eerder geneigd zijn om 'opname onbekend' in te vullen, waar anderen 'opgenomen' zouden vermelden; dit leidt volgens de in dit onderzoek toegepaste systematiek tot een hogere ophoogfactor. Wanneer er voldoende duidelijkheid is over de provinciale registratiegraad, kan ook op provinciaal niveau de ontwikkeling worden gevolgd met behulp van opgehoogde aantallen.
De ophoogfactoren variëren ook sterk naar wijze van verkeersdeelname. Wij bevelen aan nader onderzoek te doen naar de aard van de fietsongevallen (die in sterke mate ondervertegenwoordigd zijn in de VOR) om te beoordelen in welke mate het nodig en wenselijk is het verkeers-veiligheidsbeleid nadrukkelijker of anders dan nu het geval is, te richten op fietsers.

Deel C: Foutencatalogus en algemene conclusies

1. Inleiding

In dit deel worden algemene conclusies getrokken uit de verkregen gegevens. Het gaat om de volgende onderwerpen:

- De foutencatalogus, waarin de in de verschillende stadia van het onderzoek gebleken onvolkomenheden van de gekoppelde bestanden worden besproken.
- De koppelmethode; in hoeverre is deze overdraagbaar op andere combinaties van bestanden?
- De koppelresultaten; zijn er aanvullende gegevens nodig die nog niet onderzocht zijn?
- De meerwaarde van de LMR, zowel op zichzelf genomen als na koppeling.
- Wensen en aanbevelingen betreffende de bestanden LMR en VOR.
- Aanbevelingen betreffende herhaling van de koppeling.

2. De foutencatalogus

2.1. Inleiding

Bij de opzet van dit onderzoek was ervan uitgegaan dat van de beheerders van de registraties VOR en LMR, de AVV/BG en de SIG, voldoende informatie verkregen zou kunnen worden over het voorkomen van fouten in de records van de bestanden.
Zoals in deel A al uiteengezet is bleek niet voldoende kennis over deze fouten beschikbaar te zijn om de coëfficiënten van de afstandsfunctie te kunnen bepalen. Daartoe is een additionele activiteit verricht om door middel van een handmatig gestuurde koppeling deze kennis te vermeerderen. Op deze wijze zijn voldoende gegevens verkregen en is de koppeling uitgevoerd. De analyse van de bij de koppeling verkregen bestanden heeft nog meer kennis over de kwaliteit van de verschillende kenmerken opgeleverd en daaraan is in dit deel een apart hoofdstuk gewijd. Voor de duidelijkheid worden op deze plaats de kwaliteitscriteria voor registraties herhaald, zoals die in de activiteitenbeschrijving waren opgenomen.

2.2. Kwaliteitscriteria voor registraties

2.2.1. *Inleiding*

De samenleving, en in het bijzonder beleid en onderzoek, hebben behoefte aan systematisch georganiseerde gegevens met betrekking tot belangrijke maatschappelijke probleemgebieden zoals volksgezondheid en verkeer. Daartoe worden omvangrijke gegevensbestanden op computers bijgehouden, die *registraties* genoemd worden.
In de meeste gevallen richt een registratie zich op een afgeperkt gebied van voor de gebruikers van de registratie belangrijke en verkrijgbare informatie. De voor dit onderzoek relevante registraties AVV/BG en LMR zijn hier goede voorbeelden van.
Het AVV/BG bestand berust op informatie die de politie (inclusief marechaussee) vastlegt van verkeersongevallen waarbij zij geroepen is. Veel relevante informatie wordt geregistreerd, maar om verschillende redenen veel ook niet. Om de werkdruk van de agenten niet te hoog te laten zijn vormen de geregistreerde kenmerken een compromis tussen vraag en mogelijkheid. Ook is veel belangrijke informatie niet kenbaar voor de Politie, zoals de ernst van de verwondingen.
Op vergelijkbare wijze kent het LMR bestand veel gegevens over medische behandeling maar zijn de verkeerskundige aspecten onderbelicht.
Daarom wordt geprobeerd door koppeling van beide bestanden tot een voor onderzoek en beleid waardevoller bestand te komen. Het gaat dus om het verbeteren van de bruikbaarheid en de kwaliteit van de gegevens en die zijn in dit geval een gecompliceerde functie van vele aspecten van de betrokken (en door koppeling te verkrijgen) registraties, die als volgt in kaart gebracht zou kunnen worden.

Een registratie van een verschijnsel bestaat uit *records*. Ieder record correspondeert met het eenmalige optreden, volgens de *definitie*, van het verschijnsel, bijvoorbeeld een verkeersongeval of een slachtoffer, dat we in het vervolg *gebeurtenis* zullen noemen. Ieder record bestaat uit een aantal *kenmerken* die variabelen bevatten die de gebeurtenis karakteriseren. Voor een verkeersongeval zijn dat bijvoorbeeld de datum, het type weg en het aantal slachtoffers. Bij de feitelijke registratie wordt aan ieder kenmerk van een record een *waarde* toegekend. Die waarde is één van een vastgelegde reeks mogelijkheden: de *codes*. Alle mogelijke waarden van alle kenmerken horen vastgelegd te zijn in het *codeboe*k.

2.2.2. *Aspecten*

De bruikbaarheid en de waarde van registraties als die van de verkeersonveiligheid hangen af van een aantal aspecten. Deze aspecten kunnen weer onderverdeeld worden naar gelang ze betrekking hebben op het bestand als zodanig, op complete records of op kenmerken binnen records. Bij ieder aspect horen concrete criteria ter beoordeling. Het AVV-bestand dient in het vervolg als voorbeeld.

2.2.3. *Bestandsniveau*

Volledigheid òf representativiteit
Dit aspect geeft aan in hoeverre de registratie de volle omvang van het te beschrijven verschijnsel in kaart brengt. Dit wordt beoordeeld aan de hand van de *definitie* van het verschijnsel dat geregistreerd wordt.
Idealiter omvat de registratie alle relevante gebeurtenissen en geen andere, in dit geval dus alle verkeersongevallen volgens de internationale definitie, òf hij vormt een representatieve steekproef die opgehoogd kan worden tot de populatie; van iedere gebeurtenis worden - nog steeds idealiter - alle relevante kenmerken foutloos vastgelegd. Dit algemene criterium wordt hieronder gepreciseerd.
Bij een als volledig opgezette registratie is de *registratiegraad* een belangrijk criterium. De registratiegraad is het quotiënt van het als verkeersslachtoffer geregistreerde aantal en de omvang van de te registreren populatie. Altijd zal een deel gemist worden, als dat groot is, zoals de circa 40% bij de AVV/BG registratie van ziekenhuisopnamen, zou de wel geregistreerde 60% opgevat kunnen worden als een steekproef.
Deze steekproef moet echter onderscheiden worden van een als steekproef *opgezette* registratie: hierbij wordt opzettelijk slechts een (meestal klein) deel van de populatie onderzocht, *onder waarborgen dat het resultaat representatief voor de populatie is.*
Representativiteit kan bereikt worden door trekking van een *aselecte* steekproef, waarbij ieder lid van de populatie een gelijke kans heeft getrokken te worden. Ook de trekking van een *gestratificeerde* steekproef kan leiden tot representativiteit, waarbij verschillende delen van de populatie (strata) apart in de steekproef voorkomen.
Van de eerder genoemde 'steekproef' van 60% is de mate van representativiteit pas door nader onderzoek - als dit - vast te stellen.

Een belangrijk punt is de mate waarin de definitie geoperationaliseerd kan worden. Is het voor de registrerende persoon mogelijk ondubbelzinnig vast te stellen of een gebeurtenis wel of niet onder de definitie valt? Een bekend

probleem bij de verkeersongevallenregistratie AVV/BG is de ondergrens van de ernst van het letsel: een kleine verschuiving in definitie van een schram of wond heeft grote invloed op het aantal letselongevallen.

Relevantie
Het aspect 'relevantie' drukt uit in hoeverre voor de beschrijving van het onderwerp van de registratie zinvolle gegevens als kenmerk in de registratie zijn opgenomen. Veel registraties bevatten maar een deel van de relevante kenmerken. Omdat de verschillende kenmerken ook een verschillende waarde hebben voor de gebruiker zal iedere gebruiker een eigen weging hanteren. Een ruw criterium is het *aantal* (bruikbare) kenmerken.

Tijdigheid
Een registratie beslaat altijd een bepaald tijdvak, vaak een jaar. Willen beleid en onderzoek snel kunnen reageren dan moet na het eind van het tijdvak niet teveel tijd verlopen tot de gegevens beschikbaar komen. Een belangrijk criterium is de duur van die wachttijd.

Continuïteit
Wil men door middel van een registratie de ontwikkeling van een verschijnsel door de tijd volgen dan moet de registratie regelmatig plaats vinden en niet (teveel) veranderen. Het gaat hier in principe om alle aspecten, maar problematisch is bijvoorbeeld het - vanaf een bepaald tijdstip - niet meer, of anders, registreren van bepaalde kenmerken. Criteria zijn of de registratie continu is, of intermitterend, of de definitie niet verandert, of en hoeveel kenmerken in de loop van de tijd vervallen en in hoeverre de mogelijke waarden van kenmerken veranderen; alles zoals moet blijken uit de reeks codeboeken.

Flexibiliteit
Soms in tegenspraak met het vorige aspect is het verlangen om de registratie aan te kunnen passen aan nieuwe verschijnselen of nieuwe inzichten. Zo kan het wenselijk zijn kenmerken toe te voegen of de mogelijke waarden van een kenmerk uit te breiden. Een criterium is ook of en in welke mate gebruikers van de registratie invloed hierop hebben.

Koppelbaarheid
Omdat registraties vaak niet alle relevante kenmerken bevatten is het van belang dat ze gekoppeld kunnen worden, zodat de twee records uit twee registraties die dezelfde gebeurtenis beschrijven tot één - relevanter - record samengevoegd kunnen worden. Dit kan alleen als de twee registraties voldoende gelijke kenmerken van voldoende kwaliteit bevatten om die paren records eenduidig te kunnen herkennen. Het mooist is een uniek kenmerk, zoals een persoonsnummer (bij het koppelen van gewonde personen), dat in beide registraties foutloos voorkomt. Het criterium is de aanwezigheid van zo'n uniek kenmerk of van een set standaardkenmerken die koppelen mogelijk maken.

Gedocumenteerdheid
Een belangrijk kwaliteitskenmerk van een registratie is de mate van documentatie van al de hier genoemde aspecten. In veel gevallen wordt het aan de gebruiker(s) overgelaten de registratie op veel van de hier behandelde kwaliteitsaspecten te beoordelen.

2.2.4. *Recordniveau*

Correctheid
Er horen geen records ten onrechte in de registratie terecht te komen (bijvoorbeeld ongevallen buiten de openbare weg) en ook geen gebeurtenissen gemist te worden zodat er geen record van ontstaat. Deze fouten worden *afwijkingen* genoemd. Bij een representatieve steekproef geldt hetzelfde voor de feitelijke populatie waaruit de steekproef getrokken wordt.
De correctheid drukt uit in hoeverre aan het bovenstaande voldaan is. Criteria zijn het aandeel gebeurtenissen die ten onrechte als record voorkomen (percentage *te veel*), en het aandeel dat gemist wordt (percentage *te weinig*). Naar mate die percentages groter worden, en zeker bij een representatieve steekproef is belangrijk of, en in welke mate, de afwijkingen systematisch verschillen van de (bedoelde) populatie. Dan vertonen onderzoeksresultaten op basis van de registratie namelijk vergelijkbare systematische afwijkingen.

2.2.5. *Kenmerkniveau*

Resolutie
Bij 'resolutie' gaat het om de mate van detail die door een kenmerk kan worden beschreven: is de leeftijd van een persoon in jaren gegeven of wordt de geboortedatum geregistreerd. We kunnen onderscheiden theoretische en praktische resolutie. Hoe meer waarden een kenmerk kan aannemen, hoe groter de theoretische resolutie is.
De registratie kan gezien worden als een meetproces met een bepaalde meetnauwkeurigheid, de resolutie moet in verhouding staan tot die meetnauwkeurigheid, hij moet dus ook niet veel groter zijn waardoor een te grote nauwkeurigheid gesuggereerd wordt. Ook is een zekere evenwichtigheid gewenst: de mate van detaillering moet gelijkmatig over het bereik van het kenmerk verdeeld zijn. Als van een groot aantal mogelijke waarden in de praktijk maar een klein deel gebruikt wordt is de praktische resolutie gering. Een voorbeeld van een groot verschil is de registratie van het ongevalstijdstip. Dit heeft een resolutie van één minuut, maar wordt in meer dan 90% van de gevallen (begrijpelijkerwijs) afgerond op vijftallen, tientallen of meer.

Nauwkeurigheid
Bij 'nauwkeurigheid' gaat het om de juistheid van de geregistreerde waarden van de kenmerken. Een kenmerk kan verkeerd geregistreerd zijn, we spreken dan van een *fout*, maar het kan ook als *onbekend* opgenomen zijn.
Voor de hand liggende criteria zijn het percentage foutieve waarden in combinatie met de aard van de fouten, alles afgezet tegen de resolutie en de meetnauwkeurigheid. De aard van een fout kan willekeurig zijn, bijvoorbeeld een verschrijving, we spreken dan van een random fout.
Hij kan echter ook van systematische aard zijn, bijvoorbeeld door een verkeerde interpretatie van het kenmerk. Ook het percentage met code 'onbekend' per kenmerk is zeer relevant.

2.3. Methode van het bepalen van foutkansen

In principe werkt de methode waarbij gegevens over foutief gecodeerde gegevens in records van de bestanden verkregen zijn op dezelfde manier als bij de handmatig gestuurde koppeling zoals beschreven in deel A. Groepen records uit elk der bestanden worden aan elkaar toegewezen op basis van overeenstemming van een deel van de beschikbare informatie, in de terminologie van de afstandsfunctie: een bepaalde, zo klein mogelijke, afstand wordt toegelaten. Op basis van de met die kleine afstand overeenkomende grote mate van overeenstemming is aannemelijk dat maar een klein deel van de groep ten onrechte gekoppeld is. De omvang van dit aandeel wordt geschat op basis van statistische argumenten. Optredende verschillen in kenmerk-waarden (voor zover ze vaker voorkomen dan het aandeel ten onrechte gekoppelde records) worden dan geïnterpreteerd als foutieve coderingen.

Essentieel bij het toepassen van deze methode van het bepalen van foutkansen is het kunnen beschikken over twee te koppelen bestanden die zoveel kenmerken - met voldoende resolutie - hebben, dat er sprake is van *redundantie*. Hiermee wordt bedoeld dat verreweg de meeste records al door een deel van de beschikbare informatie uniek bepaald worden. Deze zelfde eis ligt overigens ook ten grondslag aan de koppelmethode zelf. Bij de hier beschreven koppeling is een deel van de beschikbare informatie niet gebruikt. Het gaat om de gegevens betreffende de vervoerswijzen van slachtoffer en botspartner, en het al of niet overleden zijn van het slachtoffer in relatie tot de datum van overlijden. Het al of niet overleden zijn op zichzelf heeft wel een rol gespeeld bij de koppeling, door het toekennen van een afstand 50 als het slachtoffer ter plaatse was overleden (waarbij de politie nooit een ziekenhuis opgeeft) en een van 35 als het slachtoffer dezelfde dag is overleden.

2.4. Vervoerswijze

Als basis dient een deelverzameling van het gekoppelde bestand, namelijk dat deel dat de records met de hoogste mate van overeenstemming bevat. Deze zijn gekoppeld met afstand 0 en selectiviteit minstens 80.
In deel B is aannemelijk gemaakt dat in deze groep minder dan 1% ten onrechte gekoppelde records voorkomen. Ook is aannemelijk dat de politie, die bij uitstek deskundig is op dit terrein en ter plaatse een onderzoek heeft ingesteld, de vervoerswijze correct heeft aangegeven. Het ligt dus voor de hand gevonden verschillen toe te schrijven aan fouten in de LMR. Deze fouten kunnen in een aantal categorieën ingedeeld worden.
De eerste is het nog gebruiken van een oude codeerwijze. Dit doet zich voor bij de groep Verkeersongevallen met een motorvoertuig (E810-E819).
Van een aantal ziekenhuizen is bij de SIG bekend dat zij de oude codering om hen moverende redenen zijn blijven hanteren. Dat heeft voor dit onderzoek geen problemen opgeleverd omdat er bij de records van die ziekenhuizen een transpositie van oude naar nieuwe codering heeft plaatsgevonden. Dat daarmee niet alle gebruik van de oude codering is opgevangen blijkt uit het voorkomen van relatief grote aantallen op volgens de nieuwe codering niet - maar volgens de oude wel toegestane combinaties, zoals beschreven in *Tabel 6* uit deel B. Deze vormen 3,8% van het totaal. De helft hiervan komt voor rekening van slechts vier ziekenhuizen.

De tweede categorie lijkt een gevolg te zijn van definitie- en interpretatie-verschillen. Relatief veel Fiets-Voetganger combinaties komen voor, maar ook Bromfiets-Fiets en Motor-Bromfiets, in beide richtingen. In totaal gaat het om 4,8%. Ruim de helft hiervan (2,8%) komt voor rekening van bij de LMR als voetgangers gecodeerde slachtoffers die volgens de VOR fietsers waren. Dit zou verklaard kunnen worden doordat bij de LMR stilstaande fietsers als voetganger opgevat worden, wat uit de codeerinstructie zou kunnen worden opgevat. De rest van de gevallen kan verwarring zijn tussen fiets en snorfiets (technisch een bromfiets); en bromfiets (wellicht met scootermodel) en motorfiets.
De derde categorie wordt gevormd door min of meer random verdeelde combinaties van vervoerswijzen met uitzondering van die van de tweede categorie. Hier gaat het om 2,9% van het totaal.
Een ander type 'fout' hebben de Niet gespecificeerde vervoerswijzen. Met 9,2% vormen zij de grootste groep. De verdeling over de (VOR)-vervoers-wijzen is niet proportioneel, bij de auto's, bestelauto's en vrachtauto's zijn circa 13% niet gespecificeerd, bij de tweewielers en de voetgangers slechts circa 6%.

Verbetering zou hier bereikt kunnen worden als alle ziekenhuizen en LMR-codeurs consistent de nieuwe codering zouden toepassen, dan wel als bij de SIG bekend zou zijn waar de oude codering gehanteerd wordt, zodat conversie kan plaats vinden. Daarnaast zou bij de instructie van de codeurs op het belang van een verkeerstechnisch juiste invulling gewezen kunnen worden. De SIG zou onmogelijke en onwaarschijnlijke combinaties van de twee helften van de E-code kunnen terugkoppelen, zoals al gedaan wordt met andere codecombinaties.

2.5. Geboortedatum

Bij de handmatig gestuurde koppeling viel al op dat kleine verschillen tussen de geboortedata veel vaker voorkwamen dan verwacht werd en niet door onterechte koppeling van verschillende slachtoffers verklaard kon worden. Daarom is op een verschil van één positie in de geboortedatum een afstandsbijdrage van 44 gezet en op twee verschillen een van 110. Omdat verschillen in geregistreerde geboortedata in de afstandsfunctie doorwerken en zo ook invloed hebben op het al of niet koppelen van twee records moet daar bij de interpretatie van aandelen gekoppelden met verschillende geboortedata rekening mee gehouden worden. Vanwege de afstandbijtelling kunnen records met verschillende geboortedatum pas voorkomen vanaf KOPKWAL 3, waarin afstanden van 41 tot 65 gegroepeerd zijn, met een selectiviteit van minstens 40.
Ook hier wordt gekeken naar de deelverzameling met de hoogste koppelkwaliteit, KOPKWAL 3. Van de 4.359 records hadden 520 een verschil in één positie van de geboortedatum. De verschillende posities waren als volgt verdeeld:

Jaar 10	Jaar 1	Maand 10	Maand 1	Dag 10	Dag 1
21	182	10	106	64	137

Tabel 1. *Aantallen gekoppelde records binnen KOPKWAL 3 met 1 positie verschillend, naar positie (JJMMDD), 1992&1993.*

Het is aannemelijk dat het hier voornamelijk om terecht gekoppelde records gaat, zodat de vraag rijst waar de fouten ontstaan kunnen zijn. Omdat de ziekenhuizen meer tijd hebben voor het verzamelen van deze soort gegevens, en zij een financieel belang hebben bij de correcte identificatie van hun patiënten, terwijl - zeker bij ernstig gewonden - de politie daar minder tijd voor heeft, is het aannemelijk dat deze verschillen in hoofdzaak aan de VOR-kant ontstaan zullen zijn.

Geboortedatum onbekend komt bij de LMR niet voor. Bij de VOR gaat het om 1.067 van de 98.351 slachtoffers (circa 1%).

2.6. Geslacht

Omdat bij de handmatig gestuurde koppeling al gebleken was dat een niet overeenstemmend geslacht bij verder maximale overeenstemming zeer zeldzaam was, terwijl deze variabele weinig selectief is werd een afstand 90 toegekend. Dat heeft tot gevolg dat met verschillend geslacht gekoppelde records pas vanaf KOPKWAL 4 kunnen voorkomen. Daarin kan naast afstand 90 nog 100 voorkomen, waarbij het epoch-verschil meer dan drie uur, maar minder dan 24 uur mag zijn. Verder moet alles overeenstemmen. Van de 906 met KOPKWAL 4 gekoppelde records hadden 35 een verschillend geslacht, waarvan bij 28 het ziekenhuis aangaf dat het een vrouw was. Ook bij deze variabele is het aannemelijk dat het merendeel door de Politie incorrect is ingevuld.

Het geslacht is bij de LMR altijd ingevuld. Bij de VOR is bij 323 slachtoffers (0,3%) het geslacht onbekend.

2.7. Ziekenhuis

Dit gegeven is voor de politie problematisch omdat ziekenhuizen voortdurend fuseren, en vaak onder andere of oude benamingen bekend zijn, terwijl bij de LMR alleen de officiële aanduiding gebruikt wordt. Bij de koppeling is zowel aan een verschillend maar bekend ziekenhuis als aan een onbekend ziekenhuis de afstand 50 toegekend.
Binnen KOPKWAL 3 (de hoogste koppelingskwaliteit waarbinnen afstand 50 voorkomt) hadden 489 van de 4.359 records een verschillend, bekend ziekenhuis. Een analyse van de kruistabel van ziekenhuisnummers volgens beide registraties laat duidelijk zien dat het in circa 90% van de gevallen gaat om ziekenhuizen in de zelfde plaats of regio. Omdat de politie niet uit eigen waarneming het ziekenhuis van uiteindelijke opname kent zijn fouten van deze omvang begrijpelijk.

In 2.508 gevallen is door de politie geen bruikbaar (in een LMR-ziekenhuisnummer vertaalbaar) ziekenhuis gecodeerd.

2.8. Datum overlijden

Een vergelijking is gemaakt in het gekoppelde bestand tussen de eventuele overlijdensdata. Vier mogelijkheden zijn onderscheiden: 1: gelijke data, 2: ongelijke data, 3: geen overlijdensdatum in LMR, 4: geen overlijdensdatum in VOR. Mogelijkheid 3 kan inhouden dat de persoon niet is overleden, maar ook dat hij of zij na ontslag uit het ziekenhuis is overleden.

Mogelijkheid 4 kan ook optreden als een verkeersslachtoffer meer dan 30 dagen na het ongeval is overleden, volgens internationale afspraak wordt dat slachtoffer niet als verkeersdode maar als (opgenomen) gewonde geregistreerd.

Gelijk	Ongelijk	LMR niet	VOR niet	Totaal
508	43	46	148	745

Tabel 2. *Overeenstemming overlijdensdata LMR - VOR 1992&1993.*

Van de 148 die bij de VOR niet als overleden geregistreerd zijn blijken circa 70% wel binnen 30 dagen overleden te zijn. Hoewel bij bijna 70% van het totaal de data overeenstemmen zijn de verschillen toch groot. Het is aan te bevelen deze nader te onderzoeken.

2.9. Datum en tijdstip ongeval en opname

Bij deze koppeling is voor de eerste maal gebruik gemaakt van de gegevens die beide bestanden hebben over het tijdstip waarop de geregistreerde gebeurtenis heeft plaatsgevonden. Bij de vorige koppeling is dat niet gebeurd omdat aangenomen werd dat dit gegeven minder nauwkeurig werd geregistreerd vanwege het geringe belang dat eraan gehecht zou kunnen worden.
Bij dit onderzoek bleek al bij de handmatig gestuurde koppeling dat het tijdstip zeer goed geregistreerd werd en dat het de koppeling kon verbeteren. Van de datum was al bekend dat die betrouwbaar is.

Het bijzondere aan de variabelen datum en Epoch (datum en tijd gecombineerd) is dat - anders dan bij de overige variabelen - geen sprake is van wel of niet overeenstemmen van de waarden in de twee registraties. Bij foutloze registratie kan een negatief epochverschil (bij terechte koppeling) niet voorkomen, maar een groot positief epochverschil is zeer wel mogelijk.
Een complicatie wordt veroorzaakt door het feit dat de politie het tijdstip in uren en minuten opgeeft, terwijl de ziekenhuizen alleen de uren registreren. Daardoor kan ten onrechte een negatief epochverschil ontstaan, bijvoorbeeld als het ongeval is gebeurd op 16.30 uur, het slachtoffer is in 20 minuten bij het ziekenhuis, het wordt om 16.55 uur opgenomen, wat geregistreerd wordt als 16.00 uur. Om deze reden is een negatief epochverschil tot 30 minuten nog als maximale overeenstemming gerekend met afstand 0 in de afstandsfunctie.
Zoals in het rapport over de analyse van de gekoppelde bestanden is aangegeven bleken in 1993 enkele tientallen gevallen van negatieve epochverschillen in de range van een half uur tot 24 uur meer voor te komen dan in 1992. Nader onderzoek bij de SIG heeft uitgewezen dat in 1993 de codeervoorschriften in zoverre versoepeld zijn dat 'onbekend tijdstip' wordt toegestaan.
Helaas wordt dat in de database naar tijdstip '0', dat wil zeggen middernacht vertaald. Door deze vervroeging van het geregistreerde opnametijdstip ontstaan in die gevallen bijna altijd negatieve epochverschillen. Beter zou zijn een code voor 'onbekend' te hanteren waarmee bij de koppeling rekening gehouden kan worden.

3. De koppelmethode

3.1. Inleiding

De in dit onderzoek voor het eerst op VOR- en LMR-bestanden toegepaste koppelmethode heeft als bijzondere eigenschap dat gebruik gemaakt wordt van een afstandsfunctie. Ook records die in een bepaalde mate verschillen en dus een zekere afstand hebben kunnen voor koppeling in aanmerking komen. De gedachte hierachter is dat iedere registratie fouten bevat. Als alleen gekoppeld wordt bij volledige overeenstemming tussen de koppelkenmerken worden vele bij elkaar horende paren ten onrechte niet gekoppeld. Het toelaten van verschillen vergroot vanzelfsprekend de kans dat er onterecht gekoppeld wordt.
De bestanden dienen aan bepaalde voorwaarden te voldoen om zinvol te werken met deze methode. Deze kunnen onderverdeeld worden in voorwaarden die de relatie tussen de bestanden betreffen en voorwaarden voor de afzonderlijke bestanden.

3.2. Voorwaarden voor de relatie tussen de bestanden

De belangrijkste voorwaarde is dat voldoende overeenkomstige informatie per record in beide bestanden aanwezig is om toewijzing van een recordpaar (een record uit het ene bestand aan een uit het andere bestand) in de overgrote meerderheid van de gevallen uniek te maken. Preciezer geformuleerd: er dienen zoveel koppelkenmerken (variabelen die in beide bestanden voorkomen die dezelfde eigenschap van het record beschrijven) met voldoende resolutie te bestaan dat het aantal mogelijke combinaties van die kenmerken veel groter is dan het aantal records in beide bestanden.
Ook moet de verdeling van de records over die kenmerken zodanig gespreid zijn dat deze eigenschap ook geldt voor deelgebieden van de ruimte opgespannen door de koppelkenmerken.

3.3. Voorwaarden binnen ieder bestand

Bij een koppeling als deze moet het gaan om bestanden die in principe een volledige registratie van het te registreren verschijnsel beogen. Als de registratiegraad van beide bestanden te klein is, ontstaan er problemen omdat terechte koppeling alleen kan plaatsvinden binnen de doorsnede van de bestanden, terwijl onterechte koppeling met alle paren mogelijk blijft. Als van beide bestanden onafhankelijke steekproeven van 10% gebruikt zouden worden is de doorsnede maar 1% van die bij volledige bestanden. De vervuiling door onterecht gekoppelde records zou relatief veel groter zijn.
Ook vallen dan de argumenten voor de aannemelijkheid dat het aandeel onterecht gekoppelde records bij afstand 0 verwaarloosbaar is gedeeltelijk weg, omdat niet meer (bijna) alle administratieve meerlingen in de steekproef hoeven zitten.

Voor de afzonderlijke bestanden moet ook gelden dat de records zodanig gespreid zijn over de koppelruimte dat te dicht op elkaar staande records niet of zeer zeldzaam voorkomen. Te dicht moet hier in relatie gezien

worden met de grootte van de verplaatsingen die door fouten in de kenmerken veroorzaakt zijn. In het bijzonder moeten samenvallende records (administratieve meerlingen) zo min mogelijk voorkomen.

Daarnaast moet voldoende kennis voorhanden zijn over de foutkansen van de koppelvariabelen. Deze kennis is nodig om de afstandsfunctie te kunnen vaststellen. Bij dit onderzoek bleek die kennis in onvoldoende mate aanwezig te zijn. Omdat de beschikbare koppelkenmerken meer informatie bevatten dan minimaal benodigd (aan de bovengenoemde eisen werd al voldaan als een koppelkenmerk werd weggelaten) kon de benodigde kennis uit de bestanden zelf aangevuld worden. Als goede kennis over de foutkansen ontbreekt moeten dus voldoende redundante informatie via de koppelvariabelen verkregen kunnen worden om de foutkansen uit de bestanden zelf te verkrijgen.

Bij iedere koppeling is het nodig om de juistheid van het koppelen op onafhankelijke wijze te controleren. Dit kan gebeuren door - eventueel voor een steekproef uit de koppelresultaten - van gegevens gebruik te maken die niet bij de koppeling gebruikt zijn. Bij dit onderzoek bleek een extra kenmerk voorhanden, de wijze van verkeersdeelname, dat een onafhankelijke controle op de juistheid van het koppelen mogelijk maakte.

3.4. Overdraagbaarheid

Als aan bovenstaande voorwaarden is voldaan is de methode zeker te gebruiken voor het koppelen van andere bestanden. Het moet dan gaan om naar opzet volledige registraties van hetzelfde verschijnsel. Zij moeten voldoende gemeenschappelijke kenmerken van voldoende praktische resolutie bevatten om een koppelruimte te kunnen definiëren met een toegevoegde afstandsfunctie, zodat de afstand tussen de meeste recordparen die hetzelfde geval betreffen kleiner is dan de afstand tussen naaste buren uit de afzonderlijke bestanden.
De resultaten zijn des te bruikbaarder naarmate uit het koppelresultaat blijkt dat aan deze voorwaarde in sterkere mate is voldaan.

4. De meerwaarde van de LMR

Bij dit onderzoek is bevestigd dat de meerwaarde van het LMR-bestand ook letterlijk opgevat moet worden: het bevat een groot aantal tot de doelpopulatie behorende slachtoffers die niet in het VOR-bestand voorkomen. Deze groep, die voor een groot deel bestaat uit gewonde fietsers, is alleen via dit bestand voor onderzoek beschikbaar. Hoewel ook dit bestand records bevat die er niet in thuis horen en andere mist die wel tot de doelpopulatie behoren is geconstateerd dat de omvang van de standaardgroep (E-codes die verkeersongevallen op de openbare weg betreffen) goed overeen komt met de geschatte totale omvang van de doelpopulatie. Daarmee heeft dit bestand dus de positie van *primair bronbestand* voor de bepaling van de werkelijke omvang van de verkeersonveiligheid, voor zover het gaat om in ziekenhuizen opgenomen gewonden.
Voor onderverdelingen naar leeftijd, geslacht, datum ongeval is het zonder meer bruikbaar, voor de vervoerswijze is een vertaling naar de VOR-codering en voor andere verkeersaspecten is een koppeling met de VOR-bestand nodig.
Daarnaast levert het natuurlijk een unieke bijdrage tot onze kennis door de grote hoeveelheid medische gegevens die geregistreerd zijn. Dit loopt van het aantal verpleegdagen tot een gedetailleerde opsomming van medische verrichtingen die na de opname zijn gebeurd.

5. Aanbevelingen

5.1. Het VOR-bestand

Om de compleetheid te bevorderen zou gestreefd moeten worden het aantal naijlers zo veel mogelijk te beperken. Het verdient aanbeveling de gangbare praktijk, om ongevallen waaraan geen locatie toegevoegd kan worden niet te registreren, te wijzigen en een categorie 'locatie onbekend' toe te voegen.

In het algemeen is het aan te bevelen onbekende gegevens in een aparte categorie onder te brengen en ze niet een speciale, maar ook mogelijke, waarde te geven. Dit geldt bijvoorbeeld voor tijdstip ongeval dat op 00.00 uur gecodeerd wordt, hoewel in dit geval de onbekenden onderscheiden kunnen worden van werkelijk om middernacht gebeurde ongevallen doordat die als 00.01 uur gecodeerd worden.

De door de VOR-codeurs gehanteerde lijst van ziekenhuizen spoort niet met de jaarlijks door de LMR vastgestelde Instellingenlijst Gezondheidszorg. Beide gebruiken een drie-cijferige code, maar ze zijn verschillend.
Ook dekken de lijsten van namen van ziekenhuizen elkaar niet, door het voortdurende proces van fusies en naamsveranderingen. Aanbevolen wordt dat de VOR de LMR-codelijst van ziekenhuizen hanteert.

5.2. Het LMR-bestand

Hier wordt aanbevolen te bevorderen dat alle codeurs het 'bindend code-advies' opvolgen en de in 1984 gewijzigde codewijze voor de vervoerswijze van het slachtoffer gebruiken. Doordat nu een niet precies bekend deel van het LMR-bestand anders gecodeerd is wordt de interpretatie bemoeilijkt en koppeling met de VOR slechter.
De redactie van de codeerinstructies is soms onduidelijk wat kan leiden tot foutieve codering. Zo wordt een belangrijk deel van de ongevallen waarbij een fietser met een motorvoertuig (inclusief bromfiets) in botsing kwam geregistreerd onder code E826 ('Fietsongeval') wat onjuist is want bij deze groep ongevallen mag geen motorvoertuig betrokken zijn. Ze horen thuis onder E813 ('Verkeersongeval met een motorvoertuig door botsing met een ander voertuig').
Ook de systematiek van de E-codering spoort niet met de behoeftes uit de hoek van het verkeersonveiligheidsonderzoek. Zo worden 'Verkeers-ongevallen met motorvoertuigen' wel onderscheiden naar wel of niet gebeurd zijn op de openbare weg maar 'Ongevallen met andere weg-voertuigen niet.
Als het opnametijdstip onbekend is moet niet '0 uur' gecodeerd worden, maar 'onbekend'.

De standaardgroep, die zo goed mogelijk alle ongevallen moet omvatten die tot de doelpopulatie behoren, kan uitgebreid worden met de E-code E801, ('Spoorwegongeval door botsing met een ander object') waaronder fiets en tram gerekend worden. Deze code kent jaarlijks enkele slachtoffers, die alle goed koppelbaar bleken. Daarnaast kunnen E805, 806 en 807 beter niet opgenomen worden omdat ze zowel volgens hun definitie geen

verkeersongeval betreffen, als bij de koppeling praktisch geen doelpopulatie bleken te bevatten.
Om dezelfde reden wordt aanbevolen de E-codes E817, 818 en 828 voortaan niet meer in de standaardgroep op te nemen.

5.3. Herhaalde koppelingen

Hoewel met de nu geconstrueerde koppelmethodiek en afstandsfunctie een zeer goed koppelresultaat bereikt is, is er nog ruimte voor verbetering.
Een nog niet opgelost dilemma doet zich voor bij de beslissing om bij volgende koppelingen de vervoerswijze al of niet in de afstandsfunctie op te nemen. Vóór pleit dat zo een beter koppelresultaat bereikt kan worden, tegen dat dan een onafhankelijke controle op het resultaat bemoeilijkt wordt.
Zeker zou bij toekomstige koppelingen geen maximum afstand meer gehanteerd moeten worden. Dan kan immers geschat worden hoeveel terecht gekoppelde records nog onder de groep met afstand boven 200 zitten.
Het verdient ook aanbeveling te onderzoeken of een glijdende afstandsschaal gekoppeld kan worden aan het epochverschil, nu kent die sprongen bij bepaalde waarden.
De mogelijkheid van onafhankelijke controle geeft in principe de gelegenheid om iteratief een optimale afstandsfunctie te berekenen. Dit zal naar verwachting zeer rekenintensief zijn.
Omdat de ontwikkelde koppelprogrammatuur al beschikbaar is zal een koppeling over nieuwe jaarcijfers in veel kortere tijd kunnen gebeuren dan in de ontwikkelingsfase. Naar schatting zal aan VOR- en LMR-zijde een dag besteed moeten worden aan de preparatie van de bestanden terwijl de koppeling zelf een dag kost. De analyse van de zo verkregen gegevens zal wat meer tijd vergen, afhankelijk van de gestelde doelen. Een project van twee à drie mensweken lijkt voldoende voor het genereren van nieuwe ophoogfactoren.
Het verdient zeker aanbeveling de koppeling om de paar jaar te herhalen omdat het verleden geleerd heeft dat de registratiegraad daalt. Dit klemt des te meer als de registratiewijze van de politie gaat veranderen.

Literatuur

AVV/BG (1995). *Gebruikershandleiding Verkeersongevallenregistratie.* Versie 4, 01-01-93. T/m aanvulling 4 d.d. 15-01-1997, Adviesdienst Verkeer en Vervoer. Basisgegevens.

AVV/BG (1995). *Registratie Verkeersongevallen; Handleiding.* Ministerie van Verkeer en Waterstaat, Directoraat-Generaal Rijkswaterstaat, Adviesdienst Verkeer en Vervoer, Hoofdafdeling BG, Heerlen, oktober 1995 - nummer 10.

Blokpoel, A. (1990). *Registratie van verkeersgewonden in het Privé-ongevallenregistratiesysteem (PORS). Resultaten van een proef.* R-90-53. SWOV, Leidschendam.

Blokpoel, A. & Polak, P.H. (1991). *Koppeling tussen de Landelijke Medische Registratie (LMR) en de Verkeersongevallenregistratie (VOR) van in ziekenhuizen opgenomen verkeersgewonden.* R-91-79. SWOV, Leidschendam.

Classificatie van Ziekten (1980). SIG/Informatiecentrum voor de Gezondheidszorg, Utrecht, 1988 (tweede druk). Gebaseerd op: International Classification of Diseases, 9th revision, Clinical Modification (ICD-9-CM), 1978 en HIV Infection Codes Addendum (Rev. No 1), 1987.

Derriks, H. & Driessen, L. (1994). *Huidige verkeersongevallengegevens; Het topje van de ijsberg?* Adviesdienst Verkeer en Vervoer, Rotterdam/Heerlen.

Larsen, P.H. (1992). *Coverage and Validity of Police Reported Traffic Accidents.* In: Proceedings of the Conference ROAD SAFETY IN EUROPE, Berlijn, Duitsland. VTI Rapport 380A, Copenhagen.

Nauta, F.A. (1988). *Rapport proefkoppeling verkeersongevallenregistratie - landelijke medische registratie.* Stichting Informatiecentrum voor de Gezondheidszorg SIG, Utrecht.

SIG Zorginformatie (1995). *Instellingenlijst Gezondheidszorg, gebaseerd op de WCC-standaard (uniforme identificering van organisatorische eenheden voor de gezondheidszorg).* Utrecht, januari 1995.

Bijlage 1 Koppeling van records uit het AVV/BG-bestand met records uit het LMR-bestand

Koppeling van records uit het AVV/BG-bestand met records uit het LMR-bestand

D.H.M. Frijters
Utrecht, mei 1996
SIG Zorginformatie

SIG Zorginformatie
Postbus 14066
3508 SC Utrecht
Telefoon 030-2345611

Inhoud

1. Inleiding

Als voortzetting/uitwerking van het rapport "Registratiegraad van in ziekenhuizen opgenomen verkeersslachtoffers -Fase 2" van P.H. Polak van de Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV, bevat dit rapport een verslag van (1) de selektie van records in de twee te koppelen bestanden AVV/BG en LMR, (2) een koppeling om echte getallen te verkrijgen over een aantal belangrijke "fouten" in de twee bestanden.

De gegevens uit bovenstaande koppeling vormen naast het gebruik van de "afstandsfunctie" (zie Polak) ook een basis om de drie oorspronkelijke vragen van de opdracht tot dit onderzoek mee te beantwoorden:

- welke aanvullende informatie biedt LMR ten opzichte van de registratie door de politie?
- welke mogelijkheden biedt LMR om extra inzicht te krijgen in de kwaliteit van de door de politie verstrekte informatie?
- welke informatie geeft LMR over de volledigheid van de registratie door de politie en welke ophoogfactoren kunnen op basis daarvanworden bepaald?

2. Specificatie van de gekoppelde deelbestanden

2.1 De onderzoeksperiode

Gegevens over één kalenderjaar zijn in principe al geschikt voor de aspecten die we met de koppeling van het AVV/BG-bestand en het LMR-bestand willen onderzoeken, zie proefkoppeling van 1987. Alhoewel de computercapaciteit vanaf 1987 tot nu sterk is toegenomen, blijft het bevragen van de LMR-database en het koppelen van de twee bestanden nog steeds een omvangrijk en tijdsintensief gebeuren. In de in deze rapportage gepresenteerde gegevens is daarom volstaan met de koppeling van de gegevens uit 1993. Het te koppelen deelbestand van LMR over 1992 is al wel gereed om op basis van de afstandsfunctie te worden gekoppeld met het AVV/BG-bestand.

2.2 Uit het LMR-bestand geselecteerde records

Uit het LMR-bestand moet de -relatief kleine- deelverzameling van verkeersslachtoffers worden geselecteerd. Wat daarbij moet worden meegenomen, wat de betekenis van de verschillende codes is en waaraan verkeersslachtoffers in de LMR kunnen worden herkend is reeds eerder in het Rapport van P.H. Polak beschreven. De selectie is als volgt geweest:

Selectie-bestand #1
Over een bepaalde periode (bijv. 1993):
van alle ziekenhuizen:

- het instellingsnummer;

van de geselecteerde patiënten:

- het opnamenummer,
- het patiëntnummer,
- het geslacht,
- de datum en het uur van opname,
- de geboortedatum,
- de instelling van herkomst als ze vanuit een andere instelling zijn opgenomen,
- de opnamereden,
- de opname-urgentie,
- de leeftijd,
- het aantal verpleegdagen tot ontslag,
- de eventuele overlijdensdatum,
- de E-code op basis waarvan ze zijn geselecteerd.

Selectie-bestand #2
Inhoud gelijk aan selectie-bestand #1, behalve dat nu niet op E-codes is geselecteerd, maar op opnamen vanwege alcoholproblemen.
Selectie #1 en Selectie #2 komen tot stand door uitgebreide queries (ca. 40 selectieregels) op de ORACLE LMR-database. Selectie #1 en #2 kunnen niet worden gecombineerd, omdat de database-structuur zich daarvoor niet leent.

Opschoningsactie #1
De selectiebestanden worden binnen een spreadsheetomgeving (MS-EXCEL) gesorteerd op instellingsnummer, opnamedatum en opnamenummer.

Selectie-bestand LMR
Aan selectie-bestand #1 worden nu de "alcohol"-diagnosecode van records van selectiebestand #2 toegevoegd telkens wanneer het om dezelfde opname van dezelfde patiënt gaat. De koppelroutine die dit uitvoert is in de macro-taal van MS-EXCEL geschreven. Dit levert het uiteindelijke Selectie-bestand LMR op, waarmee de politie-gevevens uit de ongevallenregistratie AVV/BG gaan worden gekoppeld en waarmee verder wordt geanalyseerd.

2.3 Uit het AVV/BG-bestand geselecteerde records

Van het AVV/BG-bestand worden alle records genomen, d.w.z. ook records waarbij niet is aangegeven dat een patiënt naar een ziekenhuis is vervoerd of opgenomen. Binnen de records worden een aantal gegevens als volgt geslecteerd :

Selectie-bestand #3 (=Selectiebestand VOR)
Over een bepaalde periode (bijv. 1993), per botsgeval:

- het instellingsnummer als een verkeersslachtoffer naar een bepaald ziekenhuis is vervoerd. Dit nummer wordt naar de LMR-ziekenhuiscodenummers geconverteerd;
- het VORnummer,
- het geslacht,
- de datum en het uur van plaatsvinden van het ongeval,
- de geboortedatum,
- de botspartner,
- de alcoholconsumptie,
- de ernst van het letsel,
- de wijze van deelname aan het verkeer
- de leeftijd,
- het opgenomen zijn in een ziekenhuis,
- het vervoer naar een ziekenhuis,
- de eventuele overlijdensdatum,
- het botsobject.

3. Koppelroutine en Resultaten

3.1 Doel en opzet

Primair willen we weten hoe bepaalde getallen uit de AVV/BG-registratie moeten worden opgehoogd/verlaagd. Secundair willen we voor analysedoeleinden ook één of meer bestanden ter beschikking hebben met gegevens uit de LMR en de AVV/BG die met elkaar zijn "gekoppeld".

Het beste koppelgegeven dat we in beide registraties hebben is het gegeven: "de datum van opname in het ziekenhuis (LMR) minus de datum van het ongeval (AVV/BG)" (=**de epoch**). Deze tijdstippen worden waarschijnlijk voor >99,9% in beide registraties juist gecodeerd. Als we stellen dat het tijdstip van de opname in het ziekenhuis na dat van het ongeval moet liggen en daar niet verder dan 1-2 dagen vanaf mag liggen, missen we op basis van fout coderen van deze tijdstippen in elk van de registraties vermoedelijk minder dan 20 eventueel koppelbare cases per jaar. Dit koppelgegeven is daarnaast ook behoorlijk onderscheidend, omdat je er op één dag in het jaar gemiddeld slechts ca. 70 in alle ziekenhuizen tesamen (d.w.z. ± 0,5 per ziekenhuis) verwacht aan te treffen, nl. (ca. 25.000 verkeersslachteroffer-opnamen/365). De koppeling moet daarom altijd met dit gegeven starten en in aantal gevallen zich er ook toe beperken.

Daarnaast lijkt de geboortedatum het beste koppelgegeven, alhoewel de juistheid van dit gegeven in beide registraties wellicht veel minder hoog is dan het hierboven genoemde datum-gegeven.

Geslacht is een derde --alhoewel niet erg onderscheidend-- koppelgegeven dat kan worden gebruikt. De juistheid hiervan is vermoedelijk behoorlijk groot.

Met de beschikbaarheid van deze drie koppelgegevens is een koppeling op basis van E-codes uit de LMR en ongevalsbeschrijvingen in de VOR-registratie niet nodig. Wel kan op basis van analyses de afstandsfunctie tussen deze koppelvariabelen worden berekend, waarmee de uiteindelijke vraag naar het aantal patiënten in de LMR dat verkeersslachtoffer is en als zodanig wel/dan niet in de AVV/BG is geregistreerd wellicht met nog grotere precisie kan worden berekend.

In onderstaande worden nu een aantal koppelroutines geschetst en hoe met de resultaten van deze routines kan worden berekend hoeveel LMR-records in zijn totalitiet aan VOR-records kunnen worden gekoppeld, hoeveel hiaten de AVV/BG-registratie vermoedelijk heeft en hoe groot het percentage foutcoderingen vermoedelijk is.

3.2 Routine 1, Stap 1

= Koppeling van verwijzingen uit de AVV/BG naar een bepaald ziekenhuis met de records uit het selektiebestand LMR voor dit ziekenhuis, waarbij de datum van opname minus de datum van het ongeval "0" of "1" is en de geboortedatum gelijk is. Het resultaat van deze koppeling wordt apart gezet in het koppelbestand LMR # 1 en het koppelbestand AVV/BG #1. De records die gekoppeld zijn moeten uit de oorspronkelijke selectiebestanden LMR en AVV/BG worden verwijderd voordat Stap 2 van Routine 1 kan worden uitgevoerd.

Resultaten (zie ook Bijlage)
Dit leverde 9.249 recordkoppelingen op.
Van de 26.184 records uit het selectiebestand LMR werd 35,3% van de records gekoppeld. Van de ca. 49.000 records uit het selectiebestand AVV/BG was dit ca. 18,9%.
Bij nadere analyse zou het volgende kunnen worden berekend:
Van de "echte" verkeersslachtoffercodes (LMR E-codes 812 t/m 816) zijn xxxx van de xxxxx (= xx,x) in het selectiebestand LMR gekoppeld.
Van de als in een bepaald ziekenhuis opgenomen aangeduide verkeerslachteroffers in het selectiebestand zijn xxxx van de xxxxx (=xx,x%) gekoppeld.
Uit nog wat gedetailleerdere analyse kan worden bepaald hoe vaak o.a geslacht en andere variabelen 'fout' zijn gecodeerd en kan daarmee de afstandsfunctie voor deze variabelen worden bepaald.
Uit de analyse blijkt dat 3 uur na het tijdstip van de verwijzing 84,5% is opgenomen. Van dit getal kan zou gebruik gemaakt kunnen worden om de gekoppelde records na stap 2 van de koppelroutine op te hogen tot 100%.

3.3 Routine 1, Stap 2

= Koppeling van verwijzingen uit de AVV/BG naar een bepaald ziekenhuis met de records uit het selektiebestand LMR voor dit ziekenhuis, waarbij het tijdstip van opname minus het tijdstip van het ongeval minder of gelijk aan 3 uur is en het geslacht gelijk is.

Resultaten (zie ook Bijlage)
Dit leverde ca. 1600 recordkoppelingen op. Hiervan is een groot deel "vals positief". Het aantal mogelijk goede koppelingen vanwege een fout in de geboortedatum-codering is erg moeilijk door middel van een computerprogramma boven tafel te krijgen, waarop besloten is dit handmatig uit te voeren.

Dan blijkt dat van de records uit het oorspronkelijke selectiebestand LMR nog eens 391 (=1,5%) van de LMR-records werd gekoppeld. Tesamen met Stap 1 is dat inmiddels 36,8%. Van het selectiebestand AVV/BG is dit tesamen inmiddels 19,6%. Bij nadere analyse kan nu worden berekend hoeveel van de "echte" verkeersslachtoffers, codes (LMR E-codes 812 t/m 816), in het selectiebestand LMR na deze koppelingsstap zijn gekoppeld.

.

Ook kan worden bepaald hoeveel van de als in een bepaald ziekenhuis opgenomen aangeduide verkeerslachteroffers in het selectiebestand AVV/BG zijn gekoppeld.

Alvorens naar Procedure 2 over te gaan, worden uit de Selectiebestanden LMR en AVV/BG eerst weer alle gekoppelde records verwijderd (en aan de koppelbestanden toegevoegd).

3.4 Routine 2, Stap 1

= Koppeling van ongevallen (al dan niet met [bepaalde] verwijzingen naar ziekenhuizen] uit het resterende AVV/BG -bestand met de records uit het resterende selektiebestand LMR , waarbij de datum van opname minus de datum van het ongeval "0" of "1" is en de geboortedatum gelijk is. Het resultaat van deze koppeling wordt weer toegevoegd aan het koppelbestand LMR # 1 en het koppelbestand AVV/BG #1. De records die gekoppeld zijn worden ook weer uit de oorspronkelijke selectiebestanden LMR en AVV/BG verwijderd. Met de resterende selektiebestanden AVV/BG en LMR wordt niet verder gekoppeld.

Resultaten

Dit leverde nog eens 941 recordkoppelingen op. Van de oorspronkelijke 26184 records uit het selektiebestand LMR zijn nu inmiddels in totaal 10581 records gekoppeld (=40,4%). Van de oorspronkelijke ca. 49000 records uit het selektiebestand AVV/BG is inmiddels ongeveer 21,6% gekoppeld.

3.5 Routine 2, Stap 2

= Koppeling van ongevallen (al dan niet met [bepaalde] verwijzingen naar ziekenhuizen] uit het resterende AVV/BG -bestand met de records uit het resterende selektiebestand LMR , waarbij de verwijzing naar het ziekenhuis en de opname in dit ziekenhuis, de geboortedatum en het geslacht gelijk zijn. Het resultaat van deze koppeling wordt alleen gebruikt om de variabiliteit in het "epoch" te bepalen. De exercitie is dan ook alleen uitgevoerd over de resterende ongekoppelde gegevens over Kwartaal1 van 1993.

Resultaten

Dit leverde nog eens 64 recordkoppelingen op voor Kwartaal 1. Extrapolerend naar alle kwartalen gaat het om ca. 300 extra koppelingen. Hierbij bleek dat een groot deel daarvan tot 1 dag vóór de ongevalsdatum in het ziekenhuis "was opgenomen".

3.6 Extrapolaties

Nadat Routine 2, Stap 1 is uitgevoerd kan nogmaals een analyse plaatsvinden van allerlei gegevens uit de gekoppelde records en kunnen deze vergeleken worden met de resultaten bij koppelingsstap 1 uit Routine 1. Met dat al zijn er nu voldoende gegevens uit de koppelingsstappen 1 en 2 uit Routine 1 en koppelingsstap 1 uit Routine 2 om op basis daarvan een extrapolatie uit te voeren naar:

- het werkelijk aantal koppelbare records als met de foutcoderingen wordt rekening gehouden;
- het werkelijke aantal verkeersslachtoffers dat binnen een dag na het ongeval in een ziekenhuis wordt opgenomen;
- het aantal verkeersslachtoffers dat door de politie had moeten worden gerapporteerd als zijnde in een ziekenhuis opgenomen.

4. Aansluiting op berekening afstandsfunctie en foutencatalogus

Op basis van de uitkomsten van de koppelingsstappen 1 en 2 uit Routine 1 en koppelingsstap 1 uit Routine 2 en de analyse daarvan kan nu ook :

- vrij precies de diverse foutenkansen in beide registraties worden ingeschat; en

als vervolg daarop ook

- de "afstandsfuncties" worden gedefinieerd/samengesteld, zie rapport P. Polak.

Mocht het laatste mogelijk blijken, dan zou kunnen worden gekeken of zulke afstandsfuncties ook in programmatuur kunnen worden geoperationaliseerd en of zulk een operationalisatie ook leidt tot programmatuur die generiek van aard is, d.w.z. meermalig en ook bij koppelingen tussen andere registraties bruikbaar is.

Literatuur

Registratiegraad van in ziekenhuizen opgenomen verkeersslachtoffers -Fase 2. Eerste tussenrapportage Subfase A. PH Polak, SWOV, Leidschendam, maart 1996.

Bijlage 1.

Programma-teksten

VOR-gegevens

(1) Gecomprimeerde gegevens (.arj) worden geëxpandeerd tot **ZH93.txt**

- N.B. Dit is eigenlijk een ongelukkig formaat omdat SPSS dit niet kan oppakken en het bestand derhalve ook niet in SPSS in hapklare brokken voor EXCEL kunnen worden gehakt.

(1a) In Excel worden achtereenvolgens de records uit ZH93.txt ingelezen vanaf record 1 t/m 16384, vanaf 16385 t/m etc.;

(1b) dan wordt gesorteerd op datum;

(1c) tekst DATUM UUR:MIN wordt "DATUM";

(1d) gegevens over eerste kwartaal gaan naar **VOR_KW1**, van het tweede kwartaal naar **VOR_KW2**, etc. ... **VOR_KW3** en **VOR_KW4**;

(1e) binnen deze bestanden wordt gesorteerd op instellingsnummer, ongevalsdatum, geslacht;

(1f) er worden twee velden aantoegevoegd: koppel en lmr_rec.

(1g) alle records krijgen als koppelwaarde "0".

LMR-gegevens

(2) Via UNIX-box, achtergrond-job, in SPSS geschreven **SQL-stream**, zijn de bestanden **DIAG92, DIAG93, DIAG94, ALC92, ALC93 en ALC94** aangemaakt.

(2a) Voor de koppeling van gegevens over VOR-93 worden na elkaar in SPSS DIAG93.dat en DIAG94.dat geopend m.b.v. het SPPS-syntax elementje **LMRdef.sps** dat de record-layout specificeert;

(2b) Deze bestanden worden omgezet naar **DIAG93.sav** en **DIAG94.sav**. Hier worden eerst de 800-zhnrs uit verwijderd (=Curacao). Dan wordt in DIAG93.sav gesorteerd op jaar, maand en dag, waarna records over 1 jan-93 t/m 1 april-93 in bestandje **L931.dbf** worden geplaatst. Hetzelfde vindt plaats voor bestand DIAG94.sav, waarbij records over het eerste kwartaal '93 aan L931.dbf worden toegevoegd. Evenzo worden de bestanden **L932.dbf**, **L933.dbf** en **L934.dbf** (=t/m 1 januari 1994!) gevormd;

(2c) Nu wordt L931.dbf ingelezen in EXCEL en gesaved als **L931.xls** (en **L932.xls**, **L933.xls** en **L934.xls**);

(2d) Deze bestandjes worden als volgt voor de eerste koppelslag geprepareerd (L931):

- De sheet wordt gerenamed als L931;
- De opnamedatum wordt tot echte datum gemaakt via formule (=left(E2;13)&":00").
- Bij dit en bij geboortedatum wordt "0" bij opgeteld om er een getal van te maken.
- Dan opndag, opnmd, opnjaar, opnuur deleten.
- Dan sorteren op instnr, opndatum, geslacht.
- Dan 2 nieuwe velden toevoegen: ALC en koppel.
- Koppel met "0" in alle records vullen.
- Recordvelden van VOR-bestand aan recordheadings toevoegen.
- Onder aan de kolom opndat een extra datum bijplaatsen: 3 april.

(3) ALC93.dat en ALC94.dat op dezelfde manier behandelen als DIAG93.dat en DIAG94.dat, leidend tot respektievelijk **ALC93.sav, ALC94.sav** en de dbf-bestanden **ALC931.dbf, ALC932.dbf, enz.** en de xls-bestanden **ALC931.xls, ALC932.xls, enzovoort**. Bij de laatste bestanden is het verschil met L931.xls, ewnz. in klaarmaken voor de koppeling met de E-code LMR-bestanden dat geen extra velden aan het record worden toegevoegd.

(4) Nu wordt met behulp van de macro-funktie "**Vergelijk**" op het EXCEL macro-sheet **macro.xls** ALC931.xls alcohol-records aan L931.xls records gekoppeld waar dat mogelijk is. In het bestand L931.xls levert dat in de kolom ALC een klein aantal alcohol-codes op. (N.B. Omdat dit voor de koppeling in het eerste kwartaal zo weinig hits opleverde is dit bij de andere drie kwartalen achterwege gelaten.Indien nodig kan dit achteraf alsnog worden gedaan.) **Het resultaat van deze koppeling blijft L931.xls heten.**

KOPPELROUTINE LMR-VOR

(5) De eerste stap in de koppelroutine omvat het uitvoeren van de EXCEL macro-sheet-funktie "**Koppel_S1**". Deze routine koppelt de bestanden over het eerste kwartaal VOR_KW1.xls met L931.xls. Enzovoort ook over het tweede, derde en vierde kwartaal.
(5a) Eerst moet in de funktie Koppel_S1 de namen waarin verwezen wordt naar deze bestanden op KWx en L9xy worden gezet.
(5b) Vervolgens moet in de define.names voor het LMR-bestand v.w.b. de namen DATOP, GEBDAT, INSTNR en OPNR naar het juiste sheet worden verwezen en de juiste lengte van de kolom waarnaar wordt verwezen. Bij de kolom DATOP is dit één record meer dan bij de andere records.
(5c) Idem moet voor het VOR-bestand v.w.b. de namen DATUM, DDGEBBES en Imrzhnr in define.names naar het juiste sheet en de juiste kolomlengte worden verwezen. In het laatste geval blijven de records met Imrzhnr '666' (= geen verwijzing naar een ziekenhuis) buiten beschouwing.
(5d) Na de koppeling wordt eerst gekeken of de "2"-koppelingen (=administratieve tweelingen) wel echt zijn. Soms namelijk blijkt er bij de LMR tweemaal eenzelfde record in te zitten (?). Deze worden in dat geval geschrapt en in het gekoppelde VOR-bestand wordt de "2" weer op "1" gezet.
(5e) Nu worden de resultaatbestanden gesaved als **L931S1.xls** en **V931S1.xls**.
(5f) Uit L931S1 wordt nu een extractie gedaan van alle gekoppelde records (=koppelwaarde 1 of 2). Dit extract wordt gesaved in **L931k.xls**. Dan wordt het extract 'gedeleted' en het resterend bestand wordt gesaved als **L931e.xls** (=eerste koppeling). Dit bevat enkel ongekoppelde records. Het sheet wordt gerenamed als L931e.
(5g) Idem wordt uit V931S1 geëxtraheerd en **V931k.xls** gevormd en ook **V931e.xls**.

(6) De tweede stap in de koppelroutine omvat het uitvoeren van de EXCEL macro-sheet-funktie "**Koppel_S2**". Deze routine koppelt de resultaatbestanden over het eerste kwartaal V931e.xls met L931e.xls. [Enzovoort ook over het tweede, derde en vierde kwartaal.]

(6a) Eerst moet in de funktie Koppel_S2 de namen waarin verwezen wordt naar deze bestanden op 9**xy**e worden gezet.
(6b) Vervolgens moet in de define.names voor het LMR-bestand v.w.b. de namen DATOPe, GESLAe, INSTNRe en OPNRe naar het juiste sheet worden verwezen en de juiste lengte van de kolom waarnaar wordt verwezen. Bij de kolom DATOPe is dit één record meer dan bij de andere records.
(6c) Idem moet voor het VOR-bestand v.w.b. de namen DATUMe, SEXESLe en Imrzhnre in define.names naar het juiste sheet en de juiste kolomlengte worden verwezen. In het laatste geval blijven de records met Imrzhnr '666' (= geen verwijzing naar een ziekenhuis) buiten beschouwing.
(6d) Na de koppeling wordt eerst gekeken of de "4"-koppelingen (=administratieve tweelingen) wel echt zijn. Zoniet worden ze of verwijderd of op "3" gezet (zie bij (5d).
(6e) De resultaatbestanden worden gesaved als **L931S2.xls** en **V931S2.xls**.
(6f) Uit L931S2 wordt nu een extractie gedaan van alle gekoppelde records (=koppelwaarde 3 of 4). In een hulpbestandje wordt handmatig bepaald welke records 'false positive' en welke 'echte koppelingen zijn. Van dit hulpbestandje wordt een printout gemaakt. De echt te koppelen records worde toegevoegd aan het bestand **L931k.xls,** waarna dit wordt gesaved. Dan wordt deze records uit L931S2 'gedeleted' en worden de false positive koppelingen weer ongedaan gemaakt (koppeling op "0" gezet en koppelvelden leeg gemaakt). Nadat de records gesorteerd zijn op opnamedatum, geslacht worden wordt het resterend bestand gesaved als **L931t.xls** (=tweede koppeling). Dit bevat enkel ongekoppelde records. Het sheet wordt gerenamed als L931t.
(6g) Idem wordt uit V931S2 geëxtraheerd en **V931k.xls** aangevuld en **V931t.xls** gevormd.

[N.B. Het is niet noodzakelijk om de "echte koppelingen uit Stap 2 van de koppelroutine daadwerkelijk als gekoppeld record te behandelen. Men kan de koppeling ook gebruiken om een ophoging vanwege fouten in de geboortedatum mee te bepalen en het verder laten zoals het is].

(7) De derde (en laatste) stap in de koppelroutine omvat het uitvoeren van de EXCEL macro-sheet-funktie "**Koppel_S3**". Deze routine koppelt de resultaatbestanden van Stap 2 over het eerste kwartaal V931t.xls met L931t.xls. [Enzovoort ook over het tweede, derde en vierde kwartaal.]
(7a) Eerst moet in de funktie Koppel_S3 de namen waarin verwezen wordt naar deze bestanden op 9**xy**t worden gezet.
(7b) Vervolgens moet in de define.names voor het LMR-bestand v.w.b. de namen DATOPt, GESLAt, INSTNRt, GESLAt en OPNRt naar het juiste sheet worden verwezen en de juiste lengte van de kolom waarnaar wordt verwezen. Bij de kolom DATOPt is dit één record meer dan bij de andere records.
(7c) Idem moet voor het VOR-bestand v.w.b. de namen DATUMt, SEXESLt, DDGEBESt en Imrzhnrt in define.names naar het juiste sheet en de juiste kolomlengte worden verwezen.
(7d) De resulterende koppelingen worden gesaved onder dezelfde naam als vóór deze koppelingsstap, namelijk L931t.xls en V931t.xls.

(7e) De inhoud van deze bestanden wordt in zijn geheel toegevoegd aan respektievelijk L931k.xls en V931k.xls, waarna deze bestanden worden gesaved.

(8) L931k.xls wordt opnieuw gesaved als **L931k.dbf**. Dit wordt ingelezen in SPSS en gesaved als **L93k.sav**. De andere bestanden L932k.dbf, L933k.dbf en L934k.dbf worden hiermee gemerged. L93k.sav is het bestand waarmee alle resultaten kunnen worden geanalyseerd voor wat betreft de koppeling vanuit de LMR-kant bezien.

(9) Idem voor V931k.xls, dat opnieuw wordt gesaved als **V931k.dbf**, in SPSS wordt ingelezen en uiteindelijk tot het jaarbestand **V93k.sav** leidt waarmee analyses kunnen worden uitgevoerd op de koppelingen, gezien vanuit de VOR-kant.

(10) Van de L931t.xls en V931t.xls zijn de gekoppelde records verwijderd, waarna het resultaat is gesaved onder de naam **L931d.xls**, respektievelijk **V931d.xls**. Hierop is routine **Koppel_S4** op los gelaten die geen eisen stelt aan het epoch, maar waar het instellingsnummer, de geboortedatum en het geslacht gelijk moeten zijn. Het resultaat van deze routine is onder dezelfde naam, L931d.xls, respectievelijk V931d.xls gesaved. Dit is alleen uitgevoerd voor kwartaal 1 van 1993.

Bijlage 2 Listing van het koppelingsprogramma

global.pas

```
unit global;

{ author : Stephan Westen
  create : mei 1996
  project: SWOV

  juli 1996: aanpassingen van constanten
}


interface

const
 cOnBeslist = 0;
 cGekoppeld = 1;
 cNietKoppelbaar = 2;

 { coefficienten voor de afstandsfunctie }
 c1 = 100; { epoch }
 c2 = 220; { geboortedatum }
 c3 = 90;  { geslacht }
 c4 = 50;  { ziekenhuis }
 c5 = 100; { ecode }
 c6 = 50;  { ernstl }
 phi2 : real = 0.25; { geboortedatum }
 phi3 : real = 0.5;
 max_afstand = 999;
 grens_afstand = 200;

 cLogfileName = 'SWOV_log.txt';


implementation

end.
```

koppel.pas

```
unit Koppel;

{ author : Stephan Westen
  create : mei 1996
  project: SWOV

  juli 1996:
  in afstandEpoch
    de 0.0 grens verschoven naar -0.021 (& waarde verandert)
    bij -1.0 (één dag verschil) levert ook 0.4 op (12-07-1996) ipv 1.0
  in BerekenAfstand de operator <= ipv < bij de if-statements
  in afstandECode voor 817* en 828* aangepast
  in afstandERNSTSL waarden aangepast
  in afstandGebDat een test op afwijken van twee posities toegevoegd (& waarde verandert)
  in afstandECode voor 928.9* 0.7 verandert in 0.55

  augustus 1996:
  in afstandEpoch is er een interval bijgekomen (delta < 1)

  oktober 1996:
  bug gefixed in Ecode (diagnose) interval van 820-825 ipv 820 of 825
}


interface

procedure AfstandToekenning;

procedure Koppelen;


implementation

uses dm, main, sysUtils, global, generic, dialogs;



function afstandEpoch : real;
{ deze functie kan efficienter omdat delta < -1 of delta > 3 niet mogelijk is
  ivm de range op de VOR-tabel }
 var delta : real;
begin
 with dmSWOV do
  begin
   delta:= tbLMROpnDat.asFloat - tbVORdatum.asFloat;
   if delta < -0.021 then
     if delta < -1.0 then
       result:= 1.0
     else result:= 0.4  { -1.0 <= delta < -0.021 }
   else { delta >= -0.021 }
     if delta <= 0.125 then
       result:= 0.0
     else { delta > 0.125 }
       if delta < 1 then
         result:= 0.1
       else { delta > = 1 }
         if delta <= 3.0 then
            result:= 0.4
         else result:= 1.0 { delta >= 3.0 }
  end
end;

function afstandGebDat : real;
 var
  d1, d2 : string;
  i,
  aantalFout : integer;
begin
 with dmSWOV do
  begin
   if tbVORDDGEBBES.isNUll then
    result:= phi2 { geboortedatum onbekend }
   else
    begin
```

```
      result:= 1.0;
      { indien tbLMRGebDat of tbVORDDGEBBES is Null dan lege string }
      d1:= FormatDateTime('ddmmyyyy', tbLMRgebdat.value);
      d2:= FormatDateTime('ddmmyyyy', tbVORDDGEBBES.value);
      if (length(d1) = 8) and  (length(d2) = 8) then
       begin
        aantalFout:= 0;
        { tel aantal verschillen }
        for i:=1 to 8 do
         if d1[i] <> d2[i] then inc(aantalFout);
        if aantalFout = 0 then
         result:= 0.0
        else if aantalFout = 1 then
         result:= 0.2
        else if aantalFout = 2 then
         result:= 0.5
       end
     end
   end
end;

function afstandGeslacht : real;
begin
 with dmSWOV do
 try
  if uppercase(tbVORSEXESL.asString) = 'X' then
   result:= phi3
  else
    if tbVORSEXESL.asInteger = tbLMRGESL.asInteger then
     result:= 0.0
    else result:= 1.0
 except
  on e : exception do
    begin
     result:= 1.0;
     fmMain.addLogMelding(format('afstandGeslacht fout %d, %d, %s',
        [tbLMRID.asInteger, tbVORID.asInteger , e.message]));
    end
 end;
end;

function afstandZiekenhuis : real;
begin
 with dmSWOV do
 if tbVORlmrZhnr.asInteger = tbLMRINSTnr.asInteger
  then result:= 0.0
  else result:= 1.0
end;

function afstandECode : real;
 var eersteDrie : string;
begin
 eersteDrie:= copy(dmSWOV.tbLMRDIAG.asString, 1, 3);
 if (eersteDrie = '817') or (eersteDrie = '828') then
  result:= 0.9
 else if (eersteDrie = '958') or (eersteDrie = '988') then
  result:= 0.9
 else if copy(dmSWOV.tbLMRDIAG.asString, 1, 4) = '9289' then
  result:= 0.55
 else if (eersteDrie = '820') or (eersteDrie = '821') or (eersteDrie = '822')
      or (eersteDrie = '823') or (eersteDrie = '824') or (eersteDrie = '825') then
  result:= 0.5
 else result:= 0.0;
end;

function afstandERNSTSL : real;
begin
 with dmSWOV do
  case tbVORERNSTSL.asInteger of
   1,
   8 : result:= 0.7;
   7 : result:= 0.9;
   0,
   2,
```

koppel.pas

```
   3,
   4,
   5,
   6,
   9,
   10 : result:= 0.0;
  else raise Exception.Create('Ongeldige ERNSTSL code')
 end
end;


function BerekenAfstand : real;
{ bereken de afstand voor de huidige records in tbVor en tbLMR }
{ let op: er kan een afstand uitkomen die groter is dan grens_afstand }
var
 dECode,
 dERNSTSL : real;
begin
 result:= max_afstand;
 dECode:= afstandECode;
 if (dECode*c5) <= grens_afstand then
  begin
   dERNSTSL:= afstandERNSTSL;
   if (dERNSTSL*c6) <= grens_afstand then
    begin
     result:=  (dECode * c5) +
               (dERNSTSL * c6) +
               (afstandEpoch * c1) +
               (afstandGebDat * c2) +
               (afstandGeslacht * c3) +
               (afstandZiekenhuis * c4)
    end
  end
end;


procedure StoreAfstandPointers(afstand: real);
{ update van de LMR en VOR-records }

 procedure StoreAfstandLMR;
 { De afstand wordt alleen gestored indien kleiner,
   op deze manier worden de kleinste ID's gestored }
 begin
  with dmSWOV do
  if afstand < tbLMRA2.asFloat then
   if afstand < tbLMRA1.asFLoat then
    begin
     tbLMR.Edit;
     tbLMRA2.asFloat:= tbLMRA1.asFloat;
     tbLMRA1.asFloat:= afstand;
     tbLMRP2.value := tbLMRP1.value;
     tbLMRP1.value := tbVORId.value;
     tbLMRA.asFloat:= tbLMRA1.asFloat;
     tbLMRP.value := tbLMRP1.value;
     tbLMR.Post;
    end
   else
    begin
     tbLMR.Edit;
     tbLMRA2.asFloat:= afstand;
     tbLMRP2.value := tbVORId.value;
     tbLMR.Post;
    end;
 end;

 procedure StoreAfstandVOR;
 begin
  with dmSWOV do
  if afstand < tbVORA2.asFloat then
   if afstand < tbVORA1.asFloat then
    begin
     tbVOR.Edit;
     tbVORA2.asFloat:= tbVORA1.asFloat;
     tbVORA1.asFloat:= afstand;
```

koppel.pas

```
      tbVORP2.value := tbVORP1.value;
      tbVORP1.value := tbLMRId.value;
      tbVORP.value := tbVORP1.value;
      tbVORA.asFloat:= tbVORA1.asFloat;
      tbVOR.Post;
     end
    else
     begin
      tbVOR.Edit;
      tbVORA2.asFloat:= afstand;
      tbVORP2.value := tbLMRId.value;
      tbVOR.Post;
     end;
 end;

begin
 StoreAfstandLMR;
 StoreAfstandVOR;
end;  { storeAfstandPointers }

procedure updateVoortgang;
begin
 fmMain.incVoortgang;
 if StopMatchen then
  raise TStopException.Create('Op stop-btn geklikt');
end;

procedure updateKolomK;
{ zet de records op niet-koppelbaar indien A1 > grens_afstand }
begin
 fmMain.addLogMelding('Update kolom K');
 with dmSWOV do
  begin
   tbLMR.CancelRange;
   tbVOR.CancelRange;
   tbLMR.first;
   while not(tbLMR.eof) do
    begin
     if tbLMRA1.asFloat > grens_afstand then
      begin
       tbLMR.Edit;
       tbLMRK.asInteger:= cNietKoppelbaar;
       tbLMR.Post;
      end;
     tbLMR.next;
    end;
   tbVOR.first;
   while not(tbVOR.eof) do
    begin
     if tbVORA1.asFloat > grens_afstand then
      begin
       tbVOR.Edit;
       tbVORK.asInteger:= cNietKoppelbaar;
       tbVOR.Post;
      end;
      tbVOR.next;
    end;
  end
end;

procedure AfstandToekenning;
 var
  afstand : real;
begin
 fmMain.addLogMelding('AfstandToekenning');
 with dmSWOV do
  begin
   tbLMR.indexFieldnames:= 'OPNDAT';
   tbVOR.indexFieldnames:= 'DATUM';
   { doorloop op hoogste nivo tbLMR-tabel }
   tbLMR.first;
   while not(tbLMR.eof) do
    begin
     updateVoortgang;
```

```
ppel.pas

   SetVORrange;
   tbVOR.first;
   while not tbVOR.eof do
    begin
     afstand:= berekenAfstand;
     StoreAfstandPointers(afstand);
     tbVOR.next;
    end;
   tbLMR.next;
  end;
 UpdateKolomK;
 qryLMRAantalKoppelbaar.Open;
 qryVORAantalKoppelbaar.Open;
 fmMain.addLogMelding('Aantal records koppelbaar in LMR '+qryLMRAantalKoppelbaarCOUNTOFid.a
 fmMain.addLogMelding('Aantal records koppelbaar in VOR '+qryVORAantalKoppelbaarCOUNTOFid.a
 qryVORAantalKoppelbaar.Close;
 qryVORAantalKoppelbaar.Close;
end;
d;



ocedure Koppelen;
r
ec_diepte : integer; { test-doeleinden }
antalGekoppeld : longint;

rocedure updateKoppelVoortgang;
egin
fmMain.incKoppelVoortgang(dmSWOV.tbLmr.RecordCount, dmSWOV.tbVOR.RecordCount,rec_diepte);
if StopMatchen then
 raise TStopException.Create('Op stop-btn geklikt');
nd;


rocedure probeerRecordsKoppelen(tabel1, tabel2: TTable);

procedure toewijzen(tabel1, tabel2: TTable);
 var
  t1a1, t1a2, t2a1, t2a2 : real;
  t1id, t2id : longint;
begin
 with dmSWOV do
  begin
   { zet de fields in temp. vars omdat de fields aangetast worden door de setrange }
   t1id:= tabel1.fieldByName('id').asInteger;
   t1A1:= tabel1.fieldByName('A1').asFloat;
   t1A2:= tabel1.fieldByName('A2').asFloat;
   t2id:= tabel2.fieldByName('id').asInteger;
   t2A1:= tabel2.fieldByName('A1').asFloat;
   t2A2:= tabel2.fieldByName('A2').asFloat;
   tabel1.edit;
   tabel1.fieldByName('K').asInteger := cGekoppeld;
   tabel1.fieldByName('S').asFloat := minReal(t1A2-t1A1, t2A2-t2A1);
   tabel1.fieldByName('KoppelId').asInteger := t2id;
   tabel1.post;
   tabel2.edit;
   tabel2.fieldByName('K').asInteger := cGekoppeld;
   tabel2.fieldByName('S').asFloat := minReal(t1A2-t1A1, t2A2-t2A1);
   tabel2.fieldByName('KoppelId').asInteger := t1id;
   tabel2.post;
  end;
 inc(aantalGekoppeld);
end;

procedure Tabel1_AssignP2ToP;
begin
 tabel1.edit;
 tabel1.fieldByName('p').asInteger:= tabel1.fieldByName('p2').asInteger;
 tabel1.fieldByName('a').asInteger:= tabel1.fieldByName('a2').asInteger;
 tabel1.post;
end;

procedure Tabel1_RecordWordtNietKoppelbaar;
```

koppel.pas

```
  begin
   tabel1.edit;
   tabel1.fieldByName('K').asInteger:= cNietKoppelbaar;
   tabel1.post;
  end;

 begin
  inc(rec_diepte);
  updateKoppelVoortgang;
  with dmSWOV do
   begin
    if tabel2.findkey( [cOnbeslist, tabel1.fieldByName('p').asInteger] ) then
     begin
      if tabel1.fieldByName('id').asInteger = tabel2.fieldByName('p').asInteger then
        toewijzen(tabel1, tabel2)
      else probeerRecordsKoppelen(tabel2, tabel1)
     end
    else
     begin
       Tabel1_AssignP2ToP; { probeer vervolgens voor P2 }
       if tabel2.findkey( [cOnbeslist, tabel1.fieldByName('p').asInteger] ) then
         if tabel1.fieldByName('id').asInteger = tabel2.fieldByName('p').asInteger then
           toewijzen(tabel1, tabel2)
         else probeerRecordsKoppelen(tabel2, tabel1)
       else Tabel1_RecordWordtNietKoppelbaar
     end
   end
 end; { probeerRecordsKoppelen }

 procedure VORKopierenNaarLMR;
 begin
  fmMain.addLogMelding('Kolommen kopieren');
  with dmSWOV do
   begin
    tbLMR.indexFieldnames:= 'K;ID';
    tbVOR.indexFieldnames:= 'ID';
    tbLMR.setrange([cGekoppeld], [cGekoppeld]);
    tbVOR.cancelRange;
    tbLMR.first;
    while not tbLMR.eof do
     begin
      if tbLMRK.asInteger = cGekoppeld then
       begin
        if tbVOR.findkey([tbLMRKoppelId.asInteger]) then
          begin
           tbLMR.Edit;
           tbLMRlmrzhnr.value:= tbVORlmrzhnr.value;
           tbLMRDATUM.value:= tbVORDATUM.value;
           tbLMRDDGEBBES.value:= tbVORDDGEBBES.value;
           tbLMRSEXESL.value:= tbVORSEXESL.value;
           tbLMRDDOVL.value:= tbVORDDOVL.value;
           tbLMRERNSTSL.value:= tbVORERNSTSL.value;
           tbLMRVVMK.value:= tbVORVVMK.value;
           tbLMRBOTSP.value:= tbVORBOTSP.value;
           tbLMRFUNC.value:= tbVORFUNC.value;
           tbLMROPGEN.value:= tbVOROPGEN.value;
           tbLMRVERVOER.value:= tbVORVERVOER.value;
           tbLMRALCOBE.value:= tbVORALCOBE.value;
           tbLMRLFTSL.value:= tbVORLFTSL.value;
           tbLMRVORNUM.value:= tbVORVORNUM.value;
           tbLMRZIEKHNR.value:= tbVORZIEKHNR.value;
           tbLMRKey_sla.value:= tbVORKey_sla.value;
           tbLMRx.value:= tbVORx.value;
           tbLMR.Post;
          end
        else raise exception.Create('Interne fout, vor-record niet gevonden');
       end;
      tbLMR.next;
     end
   end
 end;

begin
 fmMain.addLogMelding('Koppelen');
```

koppel.pas

```
 aantalGekoppeld:= 0;
 with dmSWOV do
  begin
    { doorlopen op rangnummer = id }
    tbLMR.indexFieldnames:= 'K;ID';
    tbVOR.indexFieldnames:= 'K;ID';
    tbLMR.setrange([cOnbeslist], [cOnbeslist]);
    tbVOR.setrange([cOnbeslist], [cOnbeslist]);
    tbLMR.first;
    while tbLMR.recordcount > 0 do
      begin
        rec_diepte:= 0;
        probeerRecordsKoppelen(tbLMR, tbVOR);
        updateKoppelVoortgang;
        tbLMR.first;
      end
  end;
 fmMain.addLogMelding('aantal gekoppeld '+intToStr(aantalGekoppeld) );
 VORKopierenNaarLMR;
end; { koppelen }


end.
```