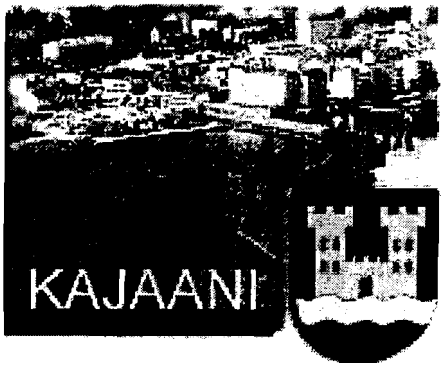
SUOMEN KUNTALIITTO
Sosiaali- ja terveysyksikkö

TERVEYDENHUOLLON 27. ATK-PÄIVÄT
4. - 5.6.2001
Sosiaali- ja terveydenhuollon tietotekniikan
ja tiedonhallinnan tutkimuksen päivät

Metodologisia näkökulmia yksilötason rekisteriaineistojen hyödyntämiseen,

Reijo Sund, Stakes



Metodologisia näkökulmia yksilötason rekisteriaineistojen hyödyntämiseen

Reijo Sund

Sosiaali- ja terveysalan tutkimus- ja kehittämiskeskus (Stakes),
Vaikuttavuuden ja oikeudenmukaisuuden tutkimusryhmä,
PL 220, 00531 Helsinki

Helsingin yliopisto,
Tilastotieteen laitos

s-posti: reijo.sund@stakes.fi

1 Johdanto

Sosiaali- ja terveydenhuollon resurssit ovat rajallisia. Näiden resurssien mahdollisimman tehokas hyödyntäminen vaatii hallinnollista suunnittelua ja poliittista tahtoa. Päätöksenteon ja valvonnan tueksi tarvitaan tieteellisesti validia ja vertailukelpoista tietoa. Suomessa terveyttä koskevat tilastot ja rekisterit ovat kansainvälisestikin katsoen varsin monipuoliset. Tietojenkäsittelyllisten resurssien kehittymisen myötä aineistojen määrä tulee kasvamaan entisestäänkin. Pelkät aineistot eivät kuitenkaan riitä - käyttökelpoisen tiedon irrottamiseksi niitä on pystyttävä myös analysoimaan monipuolisesti.

Tässä esityksessä kootaan yhteen eri tutkimusperinteistä kumpuavia - yksilötason rekisteriaineistojen hyödyntämiseen soveltuvia - metodologisia näkökulmia. Pääasiallisena tavoitteena on osoittaa, että yhdistämällä ennakkoluulottomasti sekä tilastotieteellisiä että tietojenkäsittelyllisiä ideoita myös isojen rekisteriaineistojen (esi)käsittely on teknisesti suoraviivaista. Toisaalta tarkoituksena on korostaa myös sitä, että aineistojen kasvaessa järkevien lopputuloksien saaminen edellyttää yhä enenevässä määrin sekä monipuolista metodologista osaamista että vankkaa substanssitietoutta - formalisoimalla ongelmanasettelu sopivalla tavalla saadaan eri alojen asiantuntijoiden väliselle kommunikaatiolle yhteinen kieli.

Esityksessä kuvatut ideat ovat kehittyneet "vastauksiksi" rekistereihin perustuvan terveydenhuoltotutkimuksen ja toisaalta myös tilastotuotannon todellisiin käytännön tarpeisiin. Ideoihin liittyviä menetelmiä on sovellettu esimerkiksi kirurgian potilaiden hoitoaikojen (n=75349), skitsofreenikoiden potilaspopulaation (n=95705), lonkkamurtuman jälkeisen hoidon vaikuttavuuden (n=167952), selkäleikkausten uusintariskien (n=290843) ja synnytyshoidon kustannus-vaikuttavuuden (n=692728) analysoinnissa. Yhteenvetävästi voidaan todeta, että esitettyjä ideoita on tarkoituksenmukaista hyödyntää 1) isojen aineistojen sofistikoituneeseen esikäsittelyyn, 2) substanssitiedon huomioon ottamiseen analyysiprosessissa sekä 3) ongelmanasettelun, aineiston ja mallin yhteyksien hahmottamiseen ja niiden perusteltuun yhteensovittamiseen.

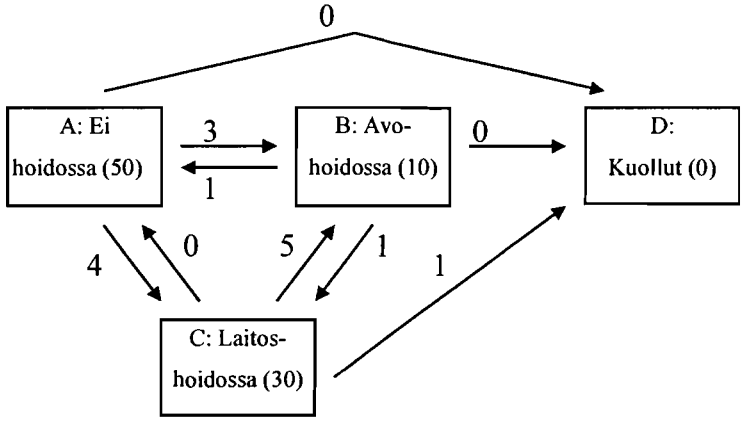
2 Tapahtumahistoria-analyysin viitekehyksestä

Dynaamisia ilmiöitä tarkasteltaessa mielenkiinto kohdistuu useiden ajassa ilmenevien tapahtumien muodostamiin tapahtumasarjoihin eli tapahtumahistorioihin. Yleensä seurataan sopivien yksiköiden liikkumista eri "tilojen" välillä; yksinkertaisimmassa tapauksessa vain yhdenlainen muutostapahtuma on mahdollinen. Monissa tilanteissa yhden tapahtuman tarkastelu ei kuitenkaan kuvaa ilmiön dynamiikkaa realistisella tavalla. Esimerkiksi hitaasti kehittyvää sairautta, joka etenee kuolemaan monien toisiaan seuraavien tilojen kautta ei pystytä kattavasti mallintamaan havainnoimalla ainoastaan ensi diagnoosista kuolemaan kuluvaa aikaa.

Tarkasteltaessa ajassa ilmeneviä tapahtumasarjoja tilastollisin menetelmin on tapana puhua tapahtumahistoria-analyysista (event history analysis). Tapahtumahistoria-analyysin viitekehys sisältää käyttökelpoisia menetelmiä monista eri tutkimusperinteistä ja tarjoaa siten mahdollisuudet erittäin monipuoliseen analyysiin[1].

Yksi usein käytetty lähtökohta dynaamisen ilmiön mallintamiseen on ilmiön toimintamekanismien hahmottaminen järjestelmäksi (system). Järjestelmä määritellään joukoksi toisiinsa liittyviä asioita tai osia, jotka toimivat yhdessä tai ovat jonkinlaisessa yhteydessä siten, että niiden voidaan ajatella muodostavan eriteltävissä olevan kokonaisuuden. Tilastollinen malli määritellään tässä yhteydessä matemaattiseksi kuvaukseksi siitä, kuinka järjestelmässä tapahtuu muutoksia.

Tämä määritelmä tarvitsee tuekseen kuvauksen järjestelmään liittyvistä olioista, ilmiöistä ja toisaalta myös rajoituksista. Periaatteessa mallin muodostaminen voidaan tulkita prosessiksi, jossa hahmotetaan ongelmanasetteluun liittyvät ilmiöt mahdollisimman yksinkertaiseksi järjestelmäksi ja esitetään, kuinka järjestelmään liittyviä tekijöitä voidaan kuvata matemaattisesti. Jos ilmiötä kuvaava järjestelmä on muodostettu järkevällä tavalla, voidaan jokaisen yksilön tapahtumahistorian tulkita olevan "polku" järjestelmän läpi. Graafisesti järjestelmä voidaan hahmottaa suunnattuna verkkona, jonka solmuilla ja kaarilla on tiloja ja transitioita kuvaavia ominaisuuksia (kuva 1).



Kuva 1. Järjestelmä suunnattuna verkkona

[1] Ks. esim. Sund, Reijo (2000): Tilastollisia menetelmiä dynaamisten potilaspopulaatioiden mallintamiseen; Tapahtumahistoria-analyysia hoitoilmoitusrekisterin skitsofreenikoille. Aiheita 26/2000. Stakes. Helsinki.

3 Tapahtumasekvenssi

Aineisto on tulkittavissa tapahtumahistoria-aineistoksi, jos siitä pystytään johtamaan havaintoja, jotka ovat muotoa $(\tau_k,\mathbf{D}_k)$, jossa τ_k on tapahtuman ilmentymisaika ja $\mathbf{D}_k$ on kyseessä olevan tapahtuman "selitys" (ja n on havaintojen määrä & $\tau_1 \leq \tau_2 \leq \ldots \leq \tau_n$ & $\tau_i < \tau_j$ ainakin yhdellä $i \neq j$ & i,j,k = 1,2,...,n).

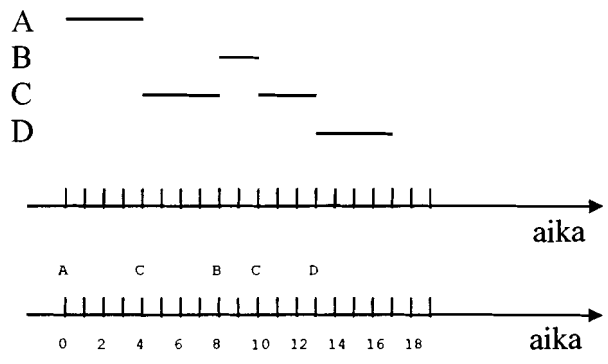
$\mathbf{D}_k$ koostuu yleensä useasta muuttujasta ja se on tulkittavissa joukoksi. Kaikki muuttujat eivät kuitenkaan välttämättä sisällä kiinnostavaa tietoa; osa tiedoista voi olla ongelmanasettelun kannalta tarpeettomia tai pääteltävissä muiden muuttujien avulla. Oleellinen tieto on yleensä tarkoituksenmukaista jakaa niin sanotun tapahtumatyypin $\mathbf{E}_k$ määrääviin ja (havaintoja sopivasti erottelevat) lisätiedot $\mathbf{i}_k$ sisältäviin muuttujien osajoukkoihin. Myöhemmin perusteltavia tarkoituksia silmälläpitäen on kätevää myös sallia tapahtuma-ajalle τ_k muunnos f_k. Jos muuta ei erikseen mainita, niin $f_k(\tau_k) = \tau_k$ (k = 1,2,...,n).

Merkitään m:llä erillisten tapahtuma-aikojen lukumäärää, t_i:llä i:nnettä erillistä tapahtuma-aikaa (i=1,2,...,m) ja tapahtumajoukolla $\mathbf{A}_i$ saman tapahtumahetken omaavien havaintojen (oleellisten tietojen) muodostamaa joukkoa eli $\mathbf{A}_i = \{(t_i,\mathbf{E}_k,\mathbf{i}_k)\}$, jossa i=1,2,...,m ja k käy joka i:llä läpi ne arvot, joilla ehto $t_i = f_k(\tau_k)$ toteutuu.

Määritellään tapahtumasekvenssi **S** (event sequence) tapahtuma-aikojen mukaan järjestetyksi tapahtumajoukkojen jonoksi eli $\mathbf{S} = < \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m >$. Olkoon lisäksi ehdollinen tapahtuma-sekvenssi eli osasekvenssi (event subsequence) jono $\mathbf{S}_\theta = < \mathbf{A}_i \mid \mathbf{A}_i \in \mathbf{S}$ ja θ toteutuu >, jossa i=1,...,m. Tilastollisen mallintamisen kannalta on huomionarvoista, että tapahtumasekvenssi on tulkittavissa merkityn pisteprosessin (marked point process) ilmentymäksi, jos kaikkien tapahtumien tapahtuma-ajat ovat erillisiä.

Edellä annettu tapahtumasekvenssin määritelmä on erittäin joustava, eikä se välttämättä tarvitse tuekseen edellisessä luvussa kuvattua tapahtumahistoria-analyysin viitekehystä. Viitekehystä hyödyntämällä tapahtumasekvenssille saadaan kuitenkin intuitiivisesti selkeä konstruktiivinen tulkinta.

t_k	$\mathbf{E}_k$
0	A
4	C
8	B
10	C
13	D



$\mathbf{S} = < \{(0,A)\}, \{(4,C)\}, \{(8,B)\}, \{(10,C)\}, \{(13,D)\} >$

Kuva 2. Esimerkki tapahtumahistoria-aineistosta

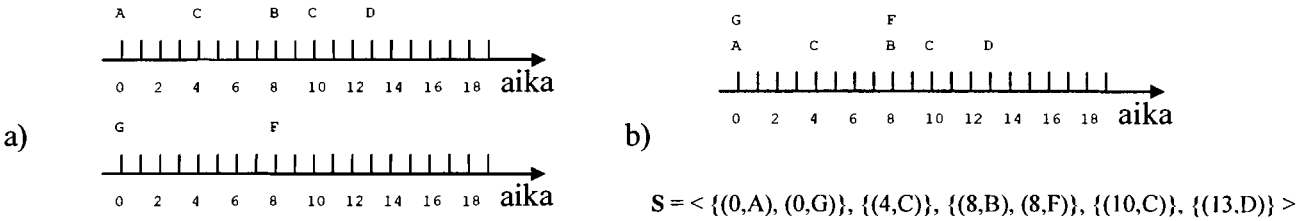
Kuvassa 2 on esimerkki tapahtumahistoria-aineistosta ja sitä vastaavasta tapahtumasekvenssistä, joka sopii kuvan 1 järjestelmän - olkoon järjestelmä nimeltään P - "tuottamaksi". Tämä yhden yksilön tapahtumahistoria on esitetty graafisesti kahdella erilaisella tavalla. Näistä esitysmuodoista ylempi havainnollistaa ehkä selkeämmin järjestelmätulkintaa, sillä siitä nähdään, että järjestelmän puitteissa kussakin tilassa viivytään kunnes siirrytään johonkin toiseen tilaan. Alemmassa tapauksessa on puolestaan merkitty näkyviin vain "kohdatut" muutostapahtumat. Koska kahden peräkkäisen tapahtuman välinen aika on tulkittavissa yhtäjaksoiseksi viipymisajaksi (length of stay) tietyssä tilassa, niin käytännössä on usein kätevää sisällyttää lisätietoihin i_k myös tätä viipymisaikaa kuvaava muuttuja, vaikka viipymisajat ovatkin erikoistapauksia lukuunottamatta helposti laskettavissa yhden yksilön tiettyyn ilmiöön liittyvän tapahtumahistorian peräkkäisistä tapahtuma-ajoista.

Tapahtumasekvenssin määritelmässä sallitaan myös useiden samanaikaisten tapahtumien ilmeneminen. Periaatteessa prosessitulkinta saadaan tässäkin tapauksessa voimaan määrittelemällä jokainen tapahtumajoukkojen yhdistelmä omaksi erilliseksi tapahtumatyypiksi. Toisin sanoen voidaan siis ajatella, että yksittäisen tapahtuman "selitys" koostuu usean saman aikaan rekisteröityneen havainnon sisältämistä tiedoista. Samanaikaiset tapahtumat on kuitenkin usein hyödyllistä pitää omina tapahtuminaan, sillä tapahtumien samanaikaisuudelle on mahdollista löytää erilaisia "luonnollisia" tulkintoja.

Kuvan 2 esimerkistä tiedetään, että kyseessä oleva aineisto kertoo miten yksilö X on liikkunut järjestelmän P puitteissa. Oletetaan, että on olemassa myös kaksitilainen suljettu järjestelmä Q, jonka tilat ovat F: "saa eläkettä" ja G: "ei saa eläkettä". Jos nyt on käytettävissä aineistoa siitä, miten yksilö X on liikkunut järjestelmässä Q, niin havaitaan sen olevan periaatteessa samaa muotoa kuin kuvassa 2 esitetty aineisto. Jos tarkastellaan aineistojen graafisia esityksiä (kuva 3 a), niin on helppo huomata, että esimerkin tapauksessa osa tapahtuma-ajoista on samoja. Tulkinnallisesti on selvää, että aineistoja "tuottavat" järjestelmät ovat rinnakkaisia (parallel), mutta se ei periaatteessa estä esittämästä molempia aineistoja yhtenä tapahtumasekvenssinä (kuva 3 b).

Rinnakkaisten järjestelmien tapauksessa on erittäin hyödyllistä sisällyttää aineistoon muuttuja, joka kertoo kummasta järjestelmästä havainto on peräisin. Rinnakkaisia järjestelmiä voi tietysti olla useampikin kuin kaksi Jos halutaan tarkastella vain tietyn järjestelmän tuottamia havaintoja, niin niihin saadaan helposti palattua ehdollisen tapahtumasekvenssin avulla eli rajoittumalla tarkastelemaan sopivaa osaa alkuperäiseen tapahtumasekvenssiin kuuluneista havainnosta.

Käytännössä tilanne on harvoin sellainen, että on havaittu vain yhden yksilön liikkumisia. Oletetaan, että pelkän yksilön X sijaan on havaittu myös yksilöiden Y ja Z liikkeet järjestelmien P ja Q puitteissa. Koko aineisto voidaan edelleen esittää yhtenä tapahtumasekvenssinä ja aineistoon on tarkoituksenmukaista lisätä myös yksilöä kuvaava muuttuja. Tällöin ehdollisia tapahtumasekvenssejä käyttäen on jälleen mahdollista rajoittua tarkastelemaan mitä tahansa osasekvenssiä.



Kuva 3. Esimerkki tilanteesta, jossa havainnoilla sama tapahtuma-aika

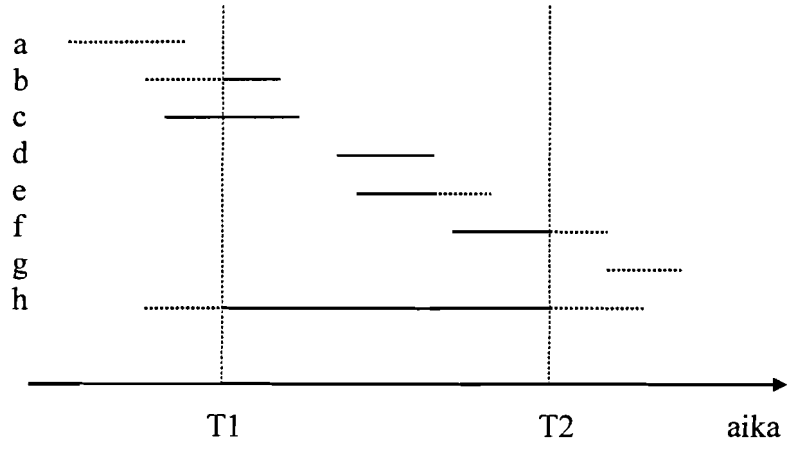
3.1 Ongelmia ja erikoistapauksia

Tähän mennessä aineiston ennakkoehtoihin ei ole erityisemmin kiinnitetty huomiota - on vain oletettu, että se on jonkin tunnetun järjestelmän tuottamaa. Käytännössä käy kuitenkin usein niin, että vaikka havaintoja tuottava järjestelmä tunnettaisiin, on havaittu tai saatavissa oleva aineisto vain kapea ikkuna taustalla olevan ilmiön käyttäytymiseen. Toisin sanoen täydellisen tapahtumasekvenssin sijaan joudutaan käytännössä aina tyytymään osasekvenssiin, joka täyttää ehdon $a < t_i < b$, jossa a ja b ovat äärellisiä vakioita ja i=1,2,...,m.

Kapeasta tarkasteluikkunasta aiheutuvia hankaluuksia on havainnollistettu kuvassa 4, jossa on kuvattu, miten yhden yksilön viipymisaika tietyssä yksittäisessä tilassa voi olla "sensuroitunut". Tapauksissa a ja g ei ole havaittu tilaan siirtymistä eikä myöskään sieltä pois siirtymistä. Tämä voi aiheuttaa suuria hankaluuksia erityisesti tilanteissa, joissa pitäisi löytää todellinen tapahtuman ensimmäinen tai viimeinen ilmentymä. Tapauksessa b tilaan siirtymistä ei ole havaittu, mutta poissiirtyminen kylläkin. Tapauksessa c sekä tilaan että sieltä pois siirtymiset on havaittu ja ongelma juontaakin juurensa siitä, että tarkasteluikkunan ulkopuolelta ei ole kattavia tietoja. Tapauksessa d on kyseessä sensuroimaton tilanne. Tapauksessa e ei ole jostain syystä havaittu "haluttua" tilasta pois siirtymistä, vaan joku "kilpaileva" poistumissyy. Tapauksessa f tilasta ei ole ehditty siirtyä pois seurannan loppuessa. Tapauksessa h tilaan ja pois siirtymiset ovat tarkasteluikkunan ulkopuolella; sensurointia voi olla monen tyyppistä.

Käytännössä kohdataan myös usein tilanteita, joissa kaikkia yksilöitä ei ole seurattu yhtä pitkään tai eri yksilöiden alkutapahtumilla on eri tapahtumahetki. Tällöin on usein tarkoituksenmukaista muuntaa aikaskaala reaaliajasta niin sanotuksi seuranta-ajaksi. Tämä onnistuu helposti valitsemalla muunnosfunktio f_k siten, että $f_k(\tau_k) = \tau_k - b_k$, jossa b_k on aina kyseessä olevan yksilön alkutapahtuman tapahtumahetki (k = 1,2,...,n).

Tarkka tapahtuma-aika ei aina ole välttämättä oleellinen tieto - pelkkä tapahtumien järjestyksen huomioonottaminen voi riittää vastaamaan tarkoituksenmukaisella tavalla joihinkin ongelman-asetteluihin. Mainittu pelkkä tapahtumien järjestyksen huomioonottava tapahtumasekvenssi saadaan valitsemalla $f_k(\tau_k) = k$ (k = 1,2,...,n). Tällaista tapahtumasekvenssiä kutsutaan tapahtuma-tyyppisekvenssiksi (event type sequence). Jos lisäksi oletetaan, että seuraavaksi ilmenevän tapahtuman todennäköisyys riippuu vain siitä tilasta, missä ollaan (Markov-ominaisuus) ja että kunkin "siirtymän" todennäköisyys on joka ajanhetkellä sama (aika-homogeenisuus), niin järjestelmässä tapahtuvia liikkeitä voidaan mallintaa Markovin ketjua käyttäen.



Kuva 4. Sensurointi

Yleisesti aikaulottuvuus sisältää kuitenkin niin paljon lisäinformaatiota ilmiöistä, että se kannattaa ehdottomasti huomioida myös malleissa. Itse asiassa useissa tapauksissa on luonnollista olettaa siirtymisen todennäköisyyden riippuvan lähtötilassa vietetystä ajasta, joten yksittäisen taajuusluvun tai volyymin sijaan on tarkoituksenmukaisempaa kuvata virtaaman ominaisuuksia lähtötilassa vietetystä ajasta riippuvan funktion avulla. Semi-Markov-malli on yksinkertaisen Markov-mallin yleistys, jossa pelkkien siirtymätodennäköisyyksien lisäksi otetaan huomioon myös tilassa ennen siirtymää vietettävä aika. Diskreetissä tapauksessa semi-Markov-malli on esitettävissä Markovin ketjujen teoriakehyksessä laajentamalla järjestelmän tila-avaruutta sopivasti.

Aina tilanne ei ole niin hyvä, että tunnettaisiin havaintoja tuottava järjestelmä tai järjestelmät. Lisäksi käytettävissä oleva aineisto saattaa sisältää runsaasti ongelmanasettelun kannalta epäoleellisia tai jopa virheellisiä havaintoja. Käytettävissä on siis mahdollisesti isokin aineisto tapahtumasekvenssin muodossa, mutta itse sekvenssin syntytavasta ei ole tiedossa kuin ehkä joitain suuntaviivoja. Tällöin tehtävänä voi olla löytää tapahtumasekvenssistä edes karkeita säännönmukaisuuksia tai todeta jonkun oletetun yhteyden olemassaolo.

Toistaiseksi kuitenkin menetelmät, joissa usean eri yksilön tapahtumahistorioista etsitään säännöllisyyksiä näyttävät saaneen osakseen yllättävän vähän huomiota. Jos esimerkiksi haluttaisiin kartoittaa mahdollisia lonkkaleikkaukseen liittyviä komplikaatioita, niin olisi järkevää valita alkutapahtumaksi lonkkaleikkaus ja sitten etsiä sitä seuraavia potilailla yleisesti ilmenneitä komplikaatioketjuja. Tässä tapauksessa komplikaation täsmällisellä tapahtumahetkellä ei välttämättä ole merkitystä, joten rajoittuminen tapahtumatyyppisekvenssien tarkasteluun voi olla järkevää.

Itse asiassa pelkkä tieto siitä, että tapahtuma on ilmennyt halutulla aikavälillä saattaa jo sisältää riittävästi informaatiota. Tällöin voi olla erittäin tarkoituksenmukaista "unohtaa" näiden havaintojen aikaulottuvuus (onnistuu muunnosfunktiolla $f_k(\tau_k) = c$, jossa c on mielivaltainen vakio ja k = 1,2,...,n). Aikaulottuvuuden hävittäminen palauttaa tapahtumasekvenssin ostoskorimalliin (market basket model). Tämä tulkinta tulee ilmeiseksi erityisesti silloin, kun tapahtumatyypit koostuvat joukosta niin sanottuja attribuutteja (attribute/item); esimerkkinä mainittakoon potilaalle yhdellä kerralla annetut diagnoosi- ja toimenpidekoodit. Tällöin on mahdollista käyttää esimerkiksi yleisten hahmojen (frequent pattern) etsimiseen soveltuvia data mining -menetelmiä yleisten tapahtuma-tyyppien (esimerkiksi diagnoosi - toimenpide-yhdistelmien) löytämiseksi.

On myös tavallista, että erilaisia tapahtumatyyppejä on niin paljon, että edellä esitetty järjestelmätulkinta "hukkuu" liialliseen yksityiskohtaisuuteen; ei välttämättä ole järkevää ajatella jokaista erilaista attribuuttijoukkoa omana tapahtumanaan. Esimerkiksi lonkkamurtuman tapauksessa ei aina olla kiinnostuneita murtuman täsmällisestä paikasta tai epäoleellisista sivudiagnooseista - tieto siitä, että kyseessä on lonkkamurtuma, saattaa riittää.

Tällaisissa tapauksissa on käytännössä hyödyllistä esikäsitellä tapahtumasekvenssiä sopivalla tavalla. Voidaan esimerkiksi hetkeksi aikaa hävittää tapahtumasekvenssin aikaulottuvuus ja tarkastella minkälaisia attribuutteja ja niiden yhdistelmistä muodostuvia tapahtumatyyppejä aineistossa ylipäätään ilmenee. Näistä attribuuteista ja tapahtumatyypeistä voidaan sitten sopivilla perusteilla (esimerkiksi "kiinnostavuus") valita tarkasteltaviksi vain osa ja toisaalta tehdä erilaisia hierarkkisia ryhmittelyitä (esimerkiksi päättää, että lonkamurtuma tarkoittaa kaikkia erikoistuneita lonkkamurtuma-diagnooseja). Kun esiprosessointi on suoritettu, niin aikaulottuvuus voidaan jälleen palauttaa ja jatkaa analysointia ongelmanasettelun kannalta tarkoituksenmukaisemmalla aineistolla.