Országos Munkaegészségügyi Intézet, MTA SZTAKI

A cluster analizis alkalmazása respiratorikus syndrómák diagnózisában

Csukás Andrásné, Mándi András, Galgóczy Gábor,
H. Gaudi István

Krónikus légzőszervi aspecifikus betegségek /KALB/ csoportjába a tüdő nem tbc-s, nem daganatos, nem keringési eredetü és nem a mellkas deformitásán alapuló megbetegedései tartoznak. Az utóbbi időben a környezeti ártalmak fokozódása miatt egyre jobban előtérbe kerül ezen betegségek epidemiológiájának vizsgálata. Idősebb korban a lakosság mintegy 30 %-a szenved ilyen tipusu betegségben és gyakori haláloki tényező is. Igy nagy jelentőségü ezen betegségek időbeni felkutatása és a kiszürt betegek megfelelő gondozása.

Az epidemiológiai vizsgálatok ismert módszerei közül a KALB felkutatására felhasználhatjuk a standardizált kérdőives módszert, a beteg orvosi vizsgálatát és a légzőszervek funkcionális állapotának müszeres ellenőrzését. A kérdőivek kitöltése a betegek bemondásán alapszik, igy az egyén szubjektivitása nagymértékben érvényesülhet, de a módszer előnye, hogy hosszabb, a felmérés előtti periódusra adhat felvilágositást. A beteg megtekintésén, valamint a kopogtatáson és hallgatódzáson alapuló orvosi vizsgálat a beteg pillanatnyi állapotáról ad képet, de az észleletek megitélése a vizsgáló orvos gyakorlatának, szaktudásának és pillanatnyi diszpozíciójának függvénye. A müszeres légzésfunkciós vizsgálatok - amennyiben módunk van a beteg kollaborációs készségétől független paraméterek mérésére - hibalehetőségként csak a

mérés pontatlanságát tartalmazzák, sok esetben ugyan csak a pillanatnyi állapotot tükrözik, de a kóros elváltozásokat bizonyitják. A fentiekből is kitünik, hogy a beteg minden szempont szerinti kielégitő felkutatása a három módszer együttes alkalmazásával várható. A jelenlegi vizsgálat célja annak eldöntése, hogy a három vizsgálati módszer által szolgáltatott adatok mennyire fedik egymást; ezáltal azt akartuk eldönteni, hogy melyik az optimális vizsgálati módszer és szükséges-e mindhárom vizsgálat együttes elvégzése.

Légzőszervi ambulanciánk 800, 20-70 év közötti férfi betegén párhuzamosan mind a három vizsgálati módszert alkalmaztuk. Ezen összehasonlitó értékelésre ez a beteganyag megfelelő volt, mert azon kivül, hogy elegendő számu és egészséges egyén is szerepel, a különböző kórformák is elegendő gyakorisággal fordulnak elő. Ezen utóbbi feltételeket egy tisztán epidemiológiai jellegü felmérésnél csak nagyon nagyszámu egyén megvizsgálása által lehetne biztositani.

Vizsgálati célkitüzésünk a légzőszervi szempontból egészséges egyének, az idült hörghurutban, /krónikus bronchitis/, hörgőszükülettel nem járó tüdőtágulatban /nonobstruktiv emphysema/ illetőlet hörgőszükülettel együtt járó tüdőtágulásban /obstruktiy emphysema/ szenvedő betegek szétválasztása volt. Ezért kérdőiven rögzitettük a köhögésre, köpetüritésre és nehézlégzésre vonatkozó panaszokat; ezek alapján 5 panaszkategóriát hoztunk létre. A fizikális vizsgálat során a tüdőtágulásra, a hörghurutra, a szív- és érrendszer betegségeire és a testalkatra vonatkozó adatokat rögzitettük; ezek alapján 9 csoportot hoztunk létre. Minden betegnél részletes testpletizmográfos, spirometriás és vérgáz analitikai vizsgálatot végeztünk a testméretek meghatározásával kiegészitve. Az igy nyert 11 funkciós érték közül a most ismertetendő

értékelési eljárásnál 4 funkciós változót használtunk fel: a léguti áramlási ellenállás /resistance/,residualis volumennek a totalis kapacitáshoz viszonyitott értéke /RV/TC %/, az artériás vér oxigén tensioja /PaO_2/ és a relativ testsuly /Broka index/. Ezek voltak azok, melyeket a beteg kollaborációs készsége nem, vagy csak alig befolyásol és csak a relativ testsulytól függőek.

Megfelelő matematikai statisztikai módszerekkel analizáltuk, hogy milyen légzésfunkciós elváltozásokkal járnak együtt az előbbi szempontok szerint kialakitott panaszkategóriák, illetve a fizikális vizsgálat során észlelt eltérések. Megállapitottuk, hogy a panaszok és a légzésfunkciós értékek közötti kapcsolat laza és csak a legsulyosabb panaszkategória esetén észleltünk kóros légzésfunkciós értékeket. A fizikális leletek és a légzésfunkciós értékek közötti kapcsolat szorosabbnak bizonyult; a sulyosabb fizikális leletek sulyosabb légzőszervi elváltozásokkal járnak együtt, mint a legsulyosabb panasz-kategória. Ezen eredményeinkről más helyen részletesebben is beszámoltunk.

Mivel a KALB epidemiológiai vizsgálata rendkivül munka- és időigényes, ezért további analiziseket végeztünk annak eldöntésére, hogy a beteg oldaláról szubjektiv kérdőives, illetve az orvos részéről szubjektiv fizikális vizsgálat adatai és az objektivnek tartható légzésfunkciós paraméterek mennyire fedik egymást. Ezek alapján kivántunk ugyanis véleményt nyilvánitani a három vizsgálati módszer értékéről. Ilyen problémák megválaszolására alkalmas módszer a cluster analizis.

Green, Frank és Robinson 1967-ben a cluster analizisre a következő definiciót adta: "a cluster analizis egy általános név, amely olyan módszerek összességére vonatkozik, mely-

nek célja, hogy azonositson hasonló dolgokat azon jellemzők-ből /minőségi vagy mennyiségi/, amelyek a dolgok tulajdonságai. Eltérően más eljárásoktól - mint például a diszkriminancia analizis - nem ismert előre, hogy mely dolgok tarotznak egy csoportba. Értéke, hogy az adatoknak egy olyan elő-csoportosítása hajtható végre, melyet az adatok természetes csoportosulása sugall".

Egy cluster analizis végrehajtása során két problémát kell megoldani; egyrészt a metrika /távolság/ definiálása, másrészt az algoritmus kiválasztása.

Legyenek $X_1, X_2, \ldots, X_n$ megfigyeléseink, amelyek kölcsönösen független és azonos eloszlásu vektor-változók a p-dimenziós térben, azaz $X_i' = (x_{i1}, x_{i2}, \ldots x_{ip})$. Esetünkben p=4. A cluster analizisben a leggyakrabban azt tételezik fel, hogy a tér Euklidesi a szokásos Euklidesi metrikával, azaz

$$\rho_E(X_i, X_j) = \sqrt{(x_i - x_j)'(x_i - x_j)} .$$

Itt feltétel, hogy a vektor komponensek függetlenek. /A kovariancia mátrix diagonális./ Nem független vektor komponensek esetén a Mahalanobis metrika a megfelelőbb

$$\rho(x_i, x_j) = \sqrt{(x_i - x_j)' V^{-1} (x_i - x_j)},$$

ahol V a kovariancia mátrix. Ha az X_i vektor komponensei különböző fontosságuak a végső következtetés szempontjából, egy megfelelően választott diagonális mátrix-szal lehet sulyozni

a komponenseket, ekkor a távolság

$$\varrho(x_i, x_j) = \sqrt{(x_i - x_j)' U' V^{-1} (x_i - x_j)} .$$

Ha a V mátrix ismeretlen, akkor helyettesithető a megfigyeléseink alapján vett becslésével:

$$\tilde{V} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' ,$$

ahol

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

A clusterek képzésére nagyon sok algoritmus ismeretes, ezek mindegyike az $X_1, X_2, \ldots, X_n$ p-dimenziós megfigyeléseinket a mintatér $S = (S_1, S_2, \ldots, S_k)$ cluster-rendszere páronként disjunkt clustereinek valamelyikébe helyezi. A cluster analizis leggyakrabban használt algoritmusai a hierarchikus algoritmusok. Ezek közül is a legismertebb a Ward-tól származó /1963/ eljárás. Az adatok csoprtositása szakaszonként történik. Az első szakaszban meg kell keresni az n számu $X_1, X_2, \ldots, X_n$ sokaság azon két elemét, melyek a választott metrika szerint legközelebb vannak egymáshoz. Ezt a két pontot középértékeikkel helyettesitjük. Ezzel n-1 pontunk maradt a p-dimenziós térben. A második lépésben ezek közül kell kiválasztani a két legközelebbit, és igy tovább. Az eljárás egy megfelelő k értékig folytatódik. Azon pontok alkotnak egy csoportot, amelyek egy cluster középpontot kialakitottak.

A hierarchikus eljárások hátránya, hogy a cluster rendszer kialakitásához az összehasonlitások igen nagy száma szükséges. Ez a szám $\sum_{l=k-2}^{n-1} l!$. Mivel a mi esetünkben n=800, igy ez az eljárás alkalmazhatatlan. Ezért az ugynevezett k-közepü eljárást használtuk. Ez k csoportból indul, amelyek mindegyike egy X_i véletlen pontból áll /i=1,2,...,k/ és ezután hozzáadja a többi X_{k+1},..., X_n pontok mindegyikét ahhoz a csoporthoz, melynek átlaga legközelebb van az adott X_i /i=k+1,...,n/ ponthoz. Az X_{k+r}-edik pontot hozzáadva valamelyik csoporthoz, ennek átlagát kiigazitjuk az X_{k+r}-edik pont figyelembevételével. Igy minden egyes szakaszban k csoport van. Egy ilyen eljárás megvalósitása k(n-k) távolság összehasonlitást igényel. Az eljárás hátránya a rögzitett k középszám. Ez feltételezi azt az előzetes információt, hogy az adatok k csoportba tömörülnek.

A k-közepü eljárás egy módositása kiküszöböli a k értékének előre való megadását, a clusterek végső száma a cluster analizis folyamata közben alakul ki. Ez a módositott algoritmus MacQueen /1967/ cikkében szerepel. A módositás lényege két előre megadott paraméter: C és R. A C az az érték, melynél közelebbi clustereket összeolvasztunk, mig azokat a pontokat, melyek minden clustertől R értéknél messzebb vannak egyik clusterhez sem vonjuk hozzá, hanem uj cluster középként kezeljük. Az algoritmus használata közben a clusterek száma hol csökken, hol növekszik, a clusterek végső számát az adatok belső, sajátos csoportosulása határozza meg.

Ezt a módszert használtuk az adataink analizisére. A program kis módositással bármilyen számu és dimenziójú adathalmazra használható. Az eljárás a C és az R értékétől függően különböző mértékben függ a kiinduló középpontok megválasztásától. Ha ugy választjuk a C és R konstansokat, hogy eljárásunk független legyen a kiinduló értékektől, akkor megnő a gépidő.

Igy előnyösebb a kettő ésszerü egyeztetése. Nálunk C=1, R=5 vagy C=1, R=4 volt. Ezért a kiindulási középpontokat mindig teoretikus alapon adtuk meg.

Az első feldolgozás során a légzésfunkciós állapot megitélésére a panasz-kategóriáknál alkalmasabbnak bizonyult fizikális lelet-kategóriáknak megfelelő középpontokból indultunk ki. Az ezek alapján megadott középpontokból kiindulva 22 csoport alakult ki, ugy, hogy az eredetileg megadott középpontok közül néhány összevonódott. A legnagyobb csoportban az esetszám 168 volt, 14 csoportban volt 10, vagy annál több megfigyelés, a többiben 5 vagy ennél kevesebb. A cluster analizis által kialakitott csoportok és a fizikális lelet csoportok között az összefüggés igen laza volt. Az egyes clusterek középpontjainak vizsgálatánál ugy tünt, hogy számos csoport létrejöttében - a főleg fiatalabb egyéneknél fennálló - átlagosnál kedvezőbb légzésfunkciós értékek játszottak szerepet. Ezért ugy gondoltuk, hogy egy transzformációt kell alkalmazni, mely a fiziológiai értékeket a fiziológiai határértékhez vonja összse. Az igy transzformált értékeket használtuk a clusterek kiépitéséhez. Az egyes légzésfunkciós paraméterket légzésfiziológiai jelentőségüknek megfelelően a távolság kiszámitásában különböző konstansokkal sulyoztuk.

A transzformációk eredményeképpen kialakult clusterek száma nem változott, de feltételezésünknek megfelelően megnőtt a legnagyobb létszámu - légzésfunkciós szempontból egészséges - csoport létszáma, csökkent a 10 vagy annál több esetet tartalmazó csoport, és az előre megadott 9 csoport közül is több olvadt be más, előre megadott középpontba. A fizikálisleletkategóriákkal való egyezés továbbra is laza maradt.

A sulyozás a clusterek kialakulását lényegesen nem befolyásolta.

Ezen eredményekből azt a következtetést vonhatjuk le, hogy a fizikális leletek alapján választott kiindulási középpontok a funkcionális paraméterek csoportosulásának belső strukturáját nem tükrözik, azaz a fizikális lelet-kategóriák és a légzésfunkciós eltérések nem fedik egymást. Tehát az adatok nem a fizikális leletek szerint válnak szét.

Az egyes cluster középpontok elemzése alapján ugy tünt, hogy a nagyobb elemszámu csoportok mindegyike különböző légzésfiziológiai syndromákat tükröz. Ennek alapján 10 uj középpontot jelöltünk ki, melyek a tüdőtágulás, a hörgőszükület, az elhizás, és a vér oxigén tensiojának különböző fokozatait veszik figyelembe. Ezen kiindulási középpontok esetében is alkalmaztuk a változók transzformációját és a sulyozást. Eredményeink a következők:

Transzformáció és sulyozás nélkül 25 cluster alakult ki, a legmagasabb elemszám 144 volt, 12 csoportnál volt az elemszám 10 vagy annál nagyobb. Valamennyi kiindulási középpontnak megfelelő cluster kialakult és valamennyiben az elemszám 10-nél nagyobb volt. Csak két jelentősebb uj csoport alakult ki 33, illetve 60-as elemszámmal.

Transzformáció alkalmazásával 26 cluster alakult ki, de ezek közül csupán 10 csoportban nagyobb az elemszám 10-nél. A legnagyobb elemszám - 301 - a fiziológiásnak itélt csoportban volt. A kiindulási középpontok által meghatározott clusterek ebben az esetben is kialakultak. A különböző sulyozások az itt elmondottakhoz képest lényeges változást nem hoztak.

A légzésfiziológiai syndromák alapján kijelölt kiindulási középpontok tehát az adatok belső strukturáját jól tükrözik, az analizis során a középpontok nem olvadtak össze, csak néhány előre nem megadott cluster alakult ki. Azonban ezek a csoportok is létező légzésfiziológiai syndromát adnak meg, melyek egymástól és az előre megadottakból élesen elkülönülnek. Ezen syndromák előfordulására a csoport közepek megadásánál nem gondoltunk.

Elméletileg a funkciós változóknak négy fokozatát különböztetjük meg: normál vagy fiziológiás, enyhén, közepesen, illetve sulyosan kóros értéket. Teoretikusan tehát 4^4 azaz 256 csoport kialakulásával kellene számolni. Ezek közül azonban egyes kombinációk nem fordulhatnak elő, mivel a vizsgált változók részben együtt járhatnak, részben következményei lehetnek egymásnak. Igy pl. egy feltünően kövér, sulyosan obstruktiv emphysemás betegnél normális vérgáz érték nem feltételezhető. Egyes csoportok pedig azért nem fordulnak elő, mert az általunk vizsgált ambuláns beteganyagban a sulyos elváltozások együttes előfordulása nem reális feltételezés, ezek a személyek már ágyhoz kötött kórházi betegek.

A légzésfiziológiai syndromák alapján kialakult nagyon jól elkülönülő és magyarázható clusterek a fizikális lelettel laza kapcsolatot mutatnak. Ennek oka lehet, hogy a fizikális eltérések nem minden esetben járnak együtt funkció károsodással, valamint a fizikális eltérések megitélése - különösen sulyosság tekintetében - igen sok szubjektiv komponenst tartalmaz. De ez nem is várható el és nem is feladata a belgyógyászati rutinvizsgálatnak. A fizikális lelet és a funkciós eltérések közötti laza kapcsolat magyarázata egyben annak, hogy a fizikális leletek alapján kialakitott kiindulási clus-

ter közepek az analizis során részben összeolvadtak, illetőleg olyan csoportok alakultak ki, melyek fizikális eltérések alapján nem különülnek el.

Vizsgálataink alapján tehát megállapithatjuk, hogy a cluster analizis alkalmas módszer és jól értékelhető felvilágositást ad többváltozós, nagytömegü adathalmazok belső kapcsolatairól. Levonhatjuk továbbá azt a következtetést is, hogy a panaszok, a fizikális lelet és a légzésfunkciós állapot felmérésére egyaránt szükség van az epidemiológiai vizsgálatokban. Ezek ugyanis nem pótolják, hanem kiegészitik egymást és csak ezek alapján lehet kialakitani egy olyan szürési rendszert, mely nemcsak a statisztikai összefüggésekre irányul, hanem lehetőleg az összes beteg embert is kiemeli.