Adatbányászati alkalmazások az egészségügyben

Vassányi István, Dr. Fogarassyné Vathy Ágnes, Tobak Tamás, Pannon Egyetem (Veszprém), Rovnyai János DSS Consulting Kft.

A klasszikusan ipari és gazdasági alkalmazásokkal rendelkező adatbányászat mára az egészségügyben is egyre nagyobb teret hódít. A cikk röviden összefoglalja az adatbányászat menetét, az egészségügyi adatok előkészítésének problémáit, majd három alkalmazási példát mutat be orvos, illetve gyógyszerforgalmi adatok elemzésére.

BEVEZETÉS

Az egészségügyi adatbányászat hátterét az adja, hogy az elektronikus páciens rekordok bevezetése óta a különféle egészségügyi információs rendszerek adatbázisaiban hatalmas mennyiségű, részben strukturált információ halmozódott fel. Az ebben rejlő tudás és összefüggések felderítése egyre kevésbé képzelhető el automatizált eszközök nélkül. Az egészségügyi adatbányászaton belül külön területet képviselnek a genetikai és molekuláris biológiai kutatások („bio-informatika"), illetve a képi adatok bányászata. A cikkben elsősorban nem ezekkel, hanem a „hagyományos" szöveges és számszerű, az ellátáshoz köthető adatok bányászatával foglalkozunk.

Az adatok elemzése során feltárt orvos-szakmai összefüggések hozzájárulhatnak a betegek „személyre szabott", eredményes ellátásához, a gazdasági jellegű összefüggések pedig az ellátás költség-hatékonyságának javításához. A cikk második felében mindkét típusú elemzésre mutatunk példát.

AZ ADATBÁNYÁSZAT FOLYAMATA

Az adatbányászat iteratív folyamat, melynek lépéseit a következőképpen foglalhatjuk össze:

- Adatelőkészítés: a rendelkezésre álló adatok beszerzése, rendszerezése, tisztítása. Ez a lépés az egész folyamat idő- és anyagi szükségletének 60-80%-át is felemésztheti. A vizsgált jelenség szempontjából leginkább lényeges változók kiválasztása például gyakran egy teljes adatbányászat lefolytatását igényli [1].
- Elemzési cél és módszer meghatározása: az adatbányászat négy alapvető módszert használ de egy gyakorlati probléma kapcsán ezek számtalan variációja és kombinációja képzelhető el (lásd a példákat a cikk második felében).
- Adatbányászat: a választott módszer (algoritmus) futtatása adatbányászatot támogató szoftver segítségével. Az eredményeket a szoftver grafikusan vagy táblázatosan közli.
- Kiértékelés és visszacsatolás: az eredmények alapján új kérdések vagy módszerek választása. Attól függően, hogy az igényelt adatok köre változik-e, visszalépés a folyamat legelejére vagy második lépéséhez.

Ahhoz, hogy az adatbányászat valóban használható eredménnyel záruljon, az adat-elemző (informatikus) szakember és az adott orvosi/gazdasági terület szakértőjének folyamatos együttműködésére van szükség [2].

EGÉSZSÉGÜGYI ADATOK ELŐKÉSZÍTÉSE

Az egészségügy a klasszikus adatbányászat egyik legnehezebb alkalmazási területe, mivel itt a más területen is jelentkező elvi és gyakorlati problémák általában halmozottan és nagyobb fokon jelentkeznek [3]. Ezeken kívül speciális, erre a területre jellemző problémák is fellépnek.

- Az adatok személyes jellege és az ebből fakadó jogi védelem megköveteli az adatrekordok anonimizálását még az adatgyűjtés megkezdése előtt.
- Különösen páciens rekordok bányászatakor igen gyakoriak a hiányzó értékek, hiszen mindenkin másmilyen vizsgálatokat végezhettek el, és a rekordok több rendszerből is származhatnak. Ezek utólagos beszerzésére nincs lehetőség, viszont az eredmény értékelhetőségéhez minél több teljes rekordra van szükség. Ezért a rekordok szűrése és a hiányzó értékek pótlása külön problémaként jelentkezik.
- Gyakoriak az adathibák, akár az információs rendszer, akár a felhasználók hibájából. Ezek felderítésére külön hihetőség-vizsgálatnak kell alávetni minden rekordot. Nagy problémát okoz, akár egy intézményen belül, a szemantikai eltérések kezelése, a szakkifejezések eltérő használata is.

Az egészségügyi adatok előkészítéséről, felmerülő nehézségekről bővebben lásd [4]. Az alábbiakban három példát mutatunk konkrét orvos-szakmai és gazdasági elemzésekre. Az első két elemzés a Pannon Egyetemen (Veszprém) készült az Intelligens Adatelemző Központ felhasználásával [5].

1. példa: ODM mérés és laboreredmények korrelációja

Az elemzés alapját az adja, hogy míg az oszteodenzitometria (ODM) egy adott mérési időpontban mutatja meg a

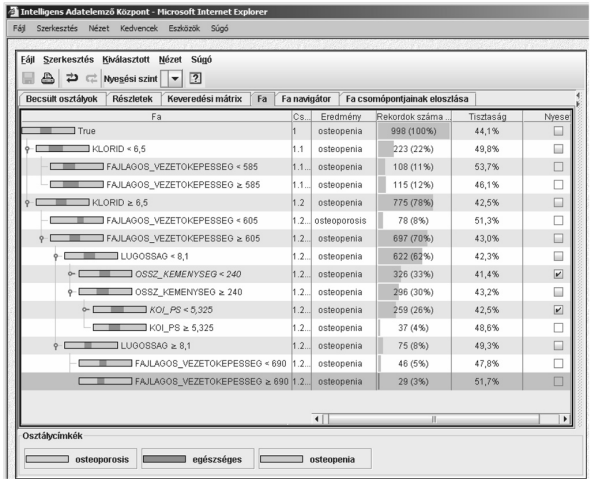
csont ásványi anyag tartalmat, a zajló bontási aktivitást a crosslaps laborvizsgálat jelzi. A crosslaps mérése egyszerű, jelentősége nagy, mert szemben az ODM 1 évével, már 3 hónap után jelzik egy antiporogén terápia hatásosságát. Elemzésünkben arra voltunk kíváncsiak, hogy a balatonfüredi DRC Kft. ellátási területén az ODM eredmény (T score) változása párhuzamosan fut-e a Crosslaps eredménnyel.

A választott módszer: korreláció-elemzés a crosslaps érték és a csontdenzitás-érték között. Az elemzés eredményeképpen igazolható volt, hogy két mérési időpont között a T-score változása szorosan korrelált a crosslaps változással. Ez azt jelenti, hogy a crosslaps vizsgálat területünkön, azaz Veszprém megyében költség-hatékony, szükséges vizsgálat.

2. példa: Az osteoporosis és az ivóvíz kapcsolata

Ebben az elemzésben arra voltunk kíváncsiak, hogy 13 Veszprém megyei településen az ivóvíz-összetétel mutat-e valamilyen kapcsolatot az osteoporosis, osteopenia előfordulásával.

A választott módszer: fa-osztályozás, azaz célunk olyan szabályok előállítása, ahol a szabály feltétel-részében az ivóvízre vonatkozó adatok foglalnak helyet, a következtetésben pedig a páciensnek a csontritkulásra vonatkozó státusza. A szabályoktól elvárjuk, hogy jól szétválasszák a vizsgált rekordokat. Az eredményül kapott döntési fában 14 szabály található (1. ábra). A fáról leolvashatjuk, hogy az első elágazás a klorid tartalom, a második a fajlagos vezetőképesség, a harmadik szinten pedig a lúgosság alapján történik, ezek tekinthetők tehát a legfontosabb változóknak. A szabályok „tisztasága“, azaz az osztályozás hatékonysága azonban olyan alacsony, hogy végül is megállapíthatjuk: az ivóvíz-összetétel és az osteoporosis között a vizsgált területen nincs összefüggés.



1. ábra
Döntési fa az ivóvíz változóival

3. példa: Gyógyszerforgalmi előrejelzések Bayes-hálókkal

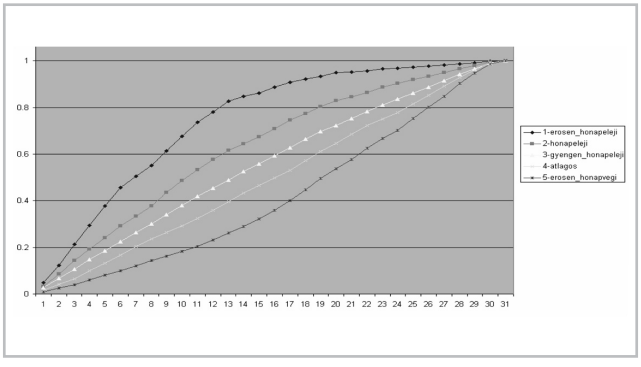
Az elemzés célja egy olyan modell kialakítása volt, melynek segítségével, egy gyógyszerforgalmazó nagykereskedelmi vállalat hó végi gyógyszer-értékesítési mennyiségei előre jelezhetővé válnak a hónap során folyamatosan, a napi adatokkal bővítve, egyre pontosabban.

A projekt megvalósítása révén lehetővé vált a vállalat számára, hogy a kialakított korszerű és hatékony információszolgáltatás, beszerzései mellett cash-flow tervezését is optimalizálja.

Ennek elérését, az egyedi üzleti és fejlesztési igények figyelembe vételével, a komplex elemzések egyszerű és könnyű alkalmazhatósága biztosította. A gyors implementáció emellett gyors megtérülést is eredményezett.

Az elemzéshez felhasznált adatok köre 571 termékről két és fél évre visszamenőlegesen, összesen 195 változóban tartalmazott információkat. Az 571 termék közül a vállalat kiválasztott 22 kiemelt terméket, melyek különböző szempontok alapján számára üzletileg a legfontosabbak voltak. Az elemzési feladat során ennek a 22 terméknek az értékesítés előrejelzéséhez kellett kialakítanunk egy hatékony előrejelző modellt.

Az elemzési feladat kulcs-mozzanata az volt, hogy megtaláljuk (kitaláljuk és előállítsuk) azokat a magyarázó változókat, amelyek a legnagyobb mértékben tudtak hozzájárulni az előrejelzési modell pontosságához. A munkafolyamat során a döntési fák (C&R Tree) és az asszociációs szabályok (GRI, apriori) mellett klaszterezést (k-központú) hajtottunk végre. Ezek segítségével tudtunk meghatározni olyan magyarázó változókat, mint például az értékesítési profil. Ez a változó normalizált formában arról nyújtott információt, hogy egy termék milyen értékesítési profillal rendelkezik egy-egy hónapban. A 2. ábrán látható grafikon a kialakított értékesítési profilokat szemlélteti:



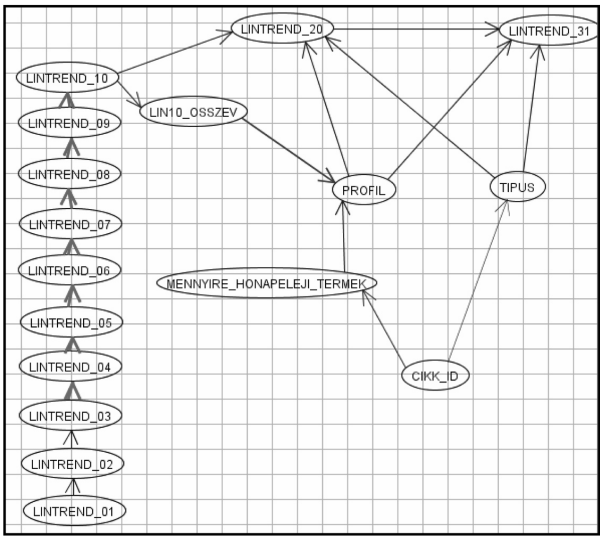
2. ábra
A kialakított értékesítési profilok

További fontos információt tartalmazott a termék típusát jelző változó. Ez a változó arra vonatkozóan adott tájékoztatást, hogy az adott termék esetében figyelembe kell-e vennünk az előrejelzés elkészítésekor éves szintű szezonalítást vagy sem.

A teljesség igénye nélkül még egy változót érdemes kiemelnünk, amely azt mutatta meg, hogy a hónap 10-ik napjáig rendelkezésünkre álló értékesítési adatok figyelembe vételével lineáris trendet követve mekkora hó végi értékesítési volumen lenne várható.

A modellépítés során partnerünk által kifejlesztett, valószínűségi hálókra épülő Bayes Generation szoftvert használtuk.

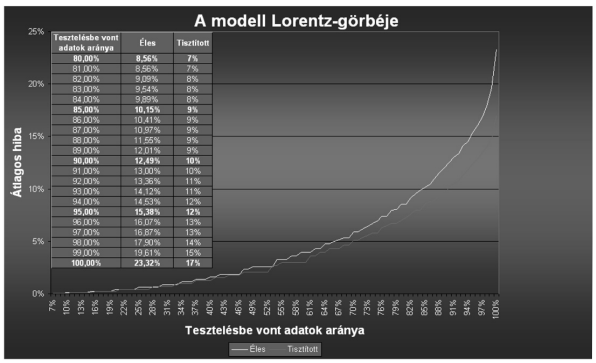
A Bayes háló vagy valószínűségi háló egy statisztikai modellt és a hozzá tartozó algoritmusok együttesét jelöli. Az algoritmusok lehetővé teszik modellek kialakítását adatokból és szakértői tudásból, illetve következtetések elvégzését az így kialakított hálókon. A valószínűségi háló véletlen csomópontokból és a közöttük lévő statisztikai függéseket reprezentáló élekből áll. Minden csomópont megfelel egy változónak, a hálózatban jelenlevő más változóktól való függéseket feltételes valószínűségi függvényekkel definiáljuk. A kialakított modell egyszerűsített szerkezetét a 3. ábra mutatja:



3. ábra
A feladat Bayes hálós modellje

A modellépítés során az elemzésbe vont adatokat véletlenszerűen három részre osztottuk. Az első rész tartalmazta a tanulóadatokat, a második a tesztadatokat, míg a harmadik a bemutatáshoz felhasznált validáló adatokat.

A kialakított modell hatékonyságát az úgynevezett Lorentz-görbével mutathatjuk be. A tesztelésbe vont adatokra vonatkozóan elkészítettük a kért előrejelzéseket, majd ezeket hasonlítottuk össze a hó végi tény adatokkal. Ezt követően növekvő sorba rendeztük a termékeket az előrejelzés százalékos hibájának nagysága szerint. A görbe vízszintesen a tesztelésbe vont termékek arányát, függőlegesen pedig ezek átlagos hibáját jelzi. Például, ha a vízszintes tengelyen 90%-hoz 12,49% tartozik (ahogy ez a lenti táblázatban szerepel is), akkor ez azt jelenti, hogy az adatok 90%-át figyelembe véve az átlagos hiba 12,49% volt.



Tesztelésbe vont adatok aránya	Éles	Tisztított
80,00%	8,56%	7%
81,00%	8,56%	7%
82,00%	9,09%	8%
83,00%	9,54%	8%
84,00%	9,89%	8%
85,00%	10,15%	9%
86,00%	10,41%	9%
87,00%	10,97%	9%
88,00%	11,55%	9%
89,00%	12,01%	9%
90,00%	12,49%	10%
91,00%	13,00%	10%
92,00%	13,36%	11%
93,00%	14,12%	11%
94,00%	14,53%	12%
95,00%	15,38%	12%
96,00%	16,07%	13%
97,00%	16,87%	13%
98,00%	17,90%	14%
99,00%	19,61%	15%
100,00%	23,32%	17%

4. ábra
A modell Lorentz-görbéje

A felső görbével jelzett adatok még tartalmazták azokat a termékeket, amelyekből csupán néhány kiszerelésnyi került értékesítésre az adott hónapban (ezért üzleti jelentőségük alacsony volt), melyeknél egy-két darabos eltérések is jelentős relatív hibát eredményezhettek. Ezeknek az elhagyásával készült az alsó görbe.

Érdemes megtekintenünk, hogy miként teljesített az előrejelző modell a kiválasztott 22 termék esetében:

Pilottermékek	Tény	10-ei előrejelzés	Hiba
1.	1144	666	-41,8%
2.	44543	37481	-15,9%
3.	30265	26122	-13,7%
4.	20117	17660	-12,2%
5.	5963	5352	-10,2%
6.	272	244	-10,2%
7.	795	725	-8,8%
8.	24172	23824	-1,4%
9.	6366	6319	-0,7%
10.	4453	4447	-0,1%
11.	36776	37559	2,1%
12.	4937	5100	3,3%
13.	3828	3983	4,0%
14.	1573	1772	12,6%
15.	1472	1777	20,7%
16.	1291	1567	21,4%
17.	7821	10250	31,1%
18.	160976	221092	37,3%
19.	6612	9457	43,0%
20.	893	1419	58,9%
21.	3596	5841	62,4%
22.	5267	14310	171,7%

1. táblázat
A teszt termékek előrejelzési eredményei

A nagy hibát okozó eseteket a sötétebb számok mutatják, ezekre végeztünk egy 20-ai becslést is, melyre a hibák jelentősen csökkentek (2. táblázat).

A nagy hibát okozó esetek között itt is található olyan, amely az abszolút kis mennyiség miatt okozott relatív nagy hibát, azaz a tisztított tesztelésbe már nem került be, illetve olyan is, amelyre egy előre nem becsülhető esemény miatt lett nagy a hiba.

Pilottermékek	Tény	10-ei előrejelzés	20-ei előrejelzés	10-ei hiba	20-ei hiba
1.	1144	666	624	-41,8%	-45,4%
18.	160976	221092	141657	37,3%	-12,0%
19.	6612	9457	4647	43,0%	-29,7%
20.	893	1419	751	58,9%	-15,9%
21.	3596	5841	3986	62,4%	10,8%
22.	5267	14310	7837	171,7%	48,8%

2. táblázat
A teszt termékek 20-ai előrejelzési eredményei

Összefoglalva megállapíthatjuk, hogy olyan modellt sikerült kialakítani, melynek segítségével, a hónap 10-ik napjáig eltelt időszak értékesítési adatai alapján átlagosan 90%-os pontossággal becsülhetjük meg egy-egy termék értékesítési volumenét (és így értékét is).

ÖSSZEFOGLALÁS

A cikk áttekintette az egészségügyi adatbányászat kihívásait és főbb problémáit, majd három alkalmazási példát mutatott az osteoporosis és a gyógyszerforgalom-előrejelzés témaköréből. További érdekes eredmények várhatóak a kialakulóban lévő országos szintű adatbázisok (pl. ESKI) adatainak feldolgozásakor.

KÖSZÖNETNYILVÁNÍTÁS

A cikkben bemutatott munkát részben az IKTA F037416 és az IKTA4-042/2001 támogatta. Ezen kívül szeretnénk megköszönni dr. Kiss József (DRC Kft, Balatonfüred) segítségét.

IRODALOMJEGYZÉK

[1] C. Baragoin et al.: Mining your own business in health care using DB2 Intelligent Miner for Data. IBM Redbook. http://www.redbooks.ibm.com/redbooks/pdfs/sg246274.pdf

[2] V. Maojo et al.: Theory, abstraction and design in medical informatics. Methods of Information in Medicine, (41):44-50, 2002.

[3] John F. Roddick et al.: Exploratory medical knowledge discovery: experiences and issues. SIGKDD Explor. Newsl., 5(1):94-99, 2003.

[4] Dr. Fogarassy György, Dr. Fogarassyné Vathy Ágnes: Egészségügyi adatok előkészítése elemzések céljából. Informatika és Menedzsment az Egészségügyben, II(8):36-41, 2003. november.

[5] A. Vathy-Fogarassy, G. Balázs, T. Tobak, I Vassányi: Intelligent Data Analysis Center: A Client/Server Mining Model over the Internet. Proc. 1st ADBIS Workshop on Data Mining and Knowledge Discovery (ADMKD'2005). Tallinn, Estonia, 15-16 September 2005, pp 57-65.

A SZERZŐK BEMUTATÁSA

Vassányi István (PhD, informatikus). 1993-ban szerzett villamosmérnöki oklevelet a Budapesti Műszaki Egyetemen. 1993-97 között a KFKI Mérés- és Számítástechnikai Kutató Intézet képfeldolgozó csoportjában programozható logikákkal dolgozott. 2000-ben szerzett informatikai PhD fokozatot a BME-n. 1997-től dolgozik a Veszprémi Egyetem Információs Rendszerek Tanszékén, jelenleg docens. Számos kutatási projekt vezetője illetve résztvevője. Kutatási területe: adatbáziskezelés, adatmodellezés, adattárházak, rendszertervezés.

Dr. Fogarassyné Vathy Ágnes (informatikus). 1995-ben szerzett matematika-fizika-számítástechnika szakos tanári diplomát a BDTF-en, majd tanulmányait a Veszprémi Egyetemen folytatta, ahol 1998-ban informatika szakos tanári diplomát szerzett. Az ELTE Informatika Doktori Iskolájának hallgatója. 1998 óta dolgozik a Veszprémi Egyetem Matematikai és Számítástechnikai Tanszékén, 2003 óta egyetemi adjunktus. Számos adatbányászattal foglalkozó szakirodalom és tudományos cikk társszerzője. Kutatási területei: adatbányászat, csoportosító algoritmusok, az adatbányászati és adattárház módszerek alkalmazása az egészségügyben, adatmodellezés, adatbázisrendszerek.

Rovnyai János (közgazdász). 2003-ban szerzett közgazdász oklevelet a Budapesti Közgazdaságtudományi és Államigazgatási Egyetem gazdaságelméleti főszakirányon. 2003 óta PhD tanulmányai mellett a DSS Consulting Kft. Adatbányászat és Intelligens Alkalmazások csoportjának munkatársa majd vezetője. Az adatbányászati üzleti projektek mellett számos kutatás-fejlesztési projekt résztvevője és vezetője is egyben. Kutatási területe: adatbányászati módszertanfejlesztés, lakossági bankok adatbányászatra épülő egyensúlyelméleti modellfejlesztése.

Tobak Tamás Jelenleg a Pannon Egyetem ötödéves műszaki informatika szakos hallgatója. Hallgatói munkája során részt vett az „Intelligens Adatelemző Központ létrehozása” című IKTA-142/2002 számú kutatási projektben. Az „Adatbányászat – a hatékonyság eszköze” című Computerbooks kiadvány, valamint több tudományos cikk társszerzője. A 2005/2006-os tanévben köztársasági ösztöndíjban részesült. Diplomadolgozatát az SAP Hungary Kft.-nél készíti, ahol az SAP Adattárház megoldásával (SAP BW) foglalkozik. Kutatási területei: adattárházak, adatbányászat, integrált vállalatirányítási rendszerek.

Folytatás a 48. oldalról

Ugyanakkor nyilvánvaló, hogy a kutatási feladatok időbeni eloszlása nagyon szabálytalan: a körülményektől függően – a napi 24 órás és heti 7 napos munkaidőt figyelembe véve – alkalmanként naponta 8-16 óra szabad kapacitása van, az ország messze legkorszerűbb MR berendezésének. Ennek az időnek a képalkotó diagnosztikában történő kihasználása elemi érdeke a magyar egészségügynek, hiszen egyrészt jelentős MR kapacitás hiány van, másrészt ez a készülék alapvető minőségi ugrást is jelent, mert egy sor olyan korszerű vizsgálati módszerre alkalmas, melyeket jelenleg Magyarországon vagy egyáltalán nem vagy csak rendkívül kis kapacitással végeznek. A készülék diagnosztikai használatbavételével javul a betegek gyógyulási esélye és egyben a betegek gyógyítására, rehabilitációjára fordítandó OEP kiadások is csökkennek.
Tekintettel a fentiekre az NKTH-tól engedélyünk van készülék diagnosztikai használatára is. Az MRKK által bevezetett új vizsgálómódszer a következők:

1. Funkcionális MR (fMRI) vizsgálatok: elsősorban a kéreg szintjén az agy különböző területeinek funkció meghatározása, mely pl.: az agytumorok, ill. a gyógyszeresen nem befolyásolható, rendkívül súlyos epilepsziák műtétjének pontos tervezéséhez nagyon fontos: a műtét okozta károsodások minimalizálhatóak ismerve a műtéti terület környezetének funkcióját. Ez a vizsgálómódszer ma még Magyarországon gyakorlatilag nem elérhető.
2. Diffuzion Tensor Immagin (DTI): segítségével a fehérállományban (agyvelő) lehet a rostok (pályák) lefutását ábrázolni s ennek segítségével például szintén a fenti műtétek tervezésében, ill. újszülöttkori károsodások terápiájában lehet lényeges javulást elérni. Eddig Magyarországon ez a vizsgálati módszer sem volt elérhető.
3. MR Spektroszkópia (MRS): az agy molekuláris szintű, in vivo, noninvasiv analizise. Segítségével az agytumorok esetén sokkal jobban elkülöníthetőek egymástól a jó-, ill. a rosszindulatú tumorok, de nagyon fontos a degeneratív betegségek, a gyulladások, a stroke, ill. az anyagcserebetegségek differenciáldiagnosztikájában is. Magyarországon ma alig végeznek ilyen vizsgálatokat.
4. Perfuzió-diffuzió vizsgálat: a stroke esetén (egy nagyobb agyi artériát vérrög zár el) a károsodás pontos felmérésére és a terápia meghatározására, melynek révén a prognózis jelentősen javulhat. Magyarországon gyakorlatilag nem végzett vizsgálat. Ráadásul az MR berendezés egy olyan klinikán van, ahol szükség esetén a vérrög oldását azonnal el tudják kezdeni (az időfaktornak a terápia eredményére döntő hatása van).

A fentieken kívül további 6-7 fontos és gyakorlatilag új vizsgálómódszert vezettek be.

Tervezik a sürgősségi indikáción alapuló cerebrovascularis MR vizsgálatok végzését (Perfiziós-diffuziós vizsgálatok stb.) is ügyeleti rendszerben: napi 24 órában és heti 7 napon, hiszen az MRKK a Neurológiai Klinikával ideális stroke szervezési és terápiás egységet alkotnának. A cerebrovasculáris betegségek gyógyítása vezető népegészségügyi feladatot jelent.
Hetente két műszakban altatásos MR vizsgálatokat terveznek gyermekeknél, mellyel az ezen a területen jelentkező hiányt jelentős fokban mérsékelni lehetne.
Természetesen végeznének hagyományos vizsgálatokat is tekintettel arra, hogy ezzel a készülékkel ezek a vizsgálatok is gyorsabban, s nagyobb diagnosztikus pontossággal végezhetőek el. Azonban céljuk, hogy elsősorban az eddig nem, vagy csak nagyon limitált számban alkalmazott korszerű vizsgálatokat végezzenek, ill. a konvencionális MR vizsgálatoknál is a minőség kerüljön az előtérbe.
A diagnosztikai vizsgálatok azért is fontosak, mert a – többek között – belőlük befolyó pénzből szeretnék a kutatásokat, a fejlesztéseket finanszírozni.
Egyenlőre nagy terveik megvalósulását erősen hátráltatja, hogy ugyan a magyar állam megteremtette az alapokat, de a működéshez forrást már nem biztosított. A takarékosság jegyében pedig az ország legmodernebb, állami kézben levő, nonprofit MR készüléke OEP finanszírozást gyakorlatilag nem kap, noha a hatásvizsgálatok nyilvánvalóan mutatják, hogy az OEP a működéshez szükséges pénz sokszorosát takaríthatná meg.