

DOKTORI (PhD) ÉRTEKEZÉS

Németh Ákos

**Ritka anyagcsere betegségek diagnosztikája
adatbányászat és gépi tanulás segítségével**

DEBRECENI EGYETEM

EGÉSZSÉGTUDOMÁNYOK DOKTORI ISKOLA

Debrecen, 2024

DOKTORI (PhD) ÉRTEKEZÉS

Ritka anyagcsere betegségek diagnosztikája adatbányászat és gépi tanulás segítségével

Németh Ákos

Témavezető: Dr. Fülöp Péter



DEBRECENI EGYETEM

EGÉSZSÉGTUDOMÁNYOK DOKTORI ISKOLA

Debrecen, 2024

Tartalomjegyzék

Tartalomjegyzék	3
Rövidítések	4
Bevezetés	6
Irodalmi áttekintés	9
Exogén és endogén lipid anyagcsere	9
Diszlipidémiák és előfordulásuk	11
Diszlipidémiák szűrése és diagnosztikája	12
Adatbányászat	17
Szöveges adatok feldolgozása	18
Bináris klasszifikáció	20
Célkitűzések	25
Anyag és módszer	26
Adattisztítás adatbányászattal	29
Összefoglaló a statisztikai elemzésekről	30
Páciensek szűrése FH és FCS diagnózisra	31
Gépi tanulás	31
Laborvizsgálatok	34
Eredmények	35
Regionális FH és FCS prevalencia	35
Javaslatok az FCS hatékonyabb diagnosztikájához	43
FH betegek komplex vizsgálata Lp(a) kategóriák szerint	47
Saját munkám	53
Új eredmények összefoglalása	54
Megbeszélés	56
Ritka betegségek prevalenciája regionális és globális összehasonlításban	56
Klinikai kritériumok és protokollok finomítása	58
FH betegek patogenezise a metabolikus eltérések és egyéb tényezők tükrében	60
A kutatás korlátai	62
Tárgyszavak	64
Összefoglalás	65
Summary	67
Felhasznált irodalom	69
Köszönetnyilvánítás	79
Függelék	80
Saját publikációk	81
Táblázatok és ábrák jegyzéke	83

Rövidítések

ACC (accuracy)	Pontosság - helyes predikciók aránya az összes predikciókhoz - klasszifikációs modellek teljesítmény metrikája
ANN (artificial neural network)	Mesterséges neurális hálózat (gépi tanulás modellek egy csoportja)
AoS (aortic stenosis)	Aortabillentyű sztenózis
Apo	Apo-lipoprotein
ASCVD (atherosclerotic cardiovascular disease)	ateroszklerotikus eredetű kardiovaszkuláris megbetegedés
BERT (bidirectional encoder representations from transformers)	Transzformer alapú, neurális hálózatra épülő nyelvi modell
CeVD (cerebrovascular disease)	Cerebrovaszkuláris megbetegedés
Chol (cholesterol)	Koleszterin
CVA (cerebrovascular accident)	Cerebrovaszkuláris esemény
CVD (cardiovascular diseases)	Kardiovaszkuláris megbetegedés
DEKK	Debreceni Egyetem Klinikai Központ
DLCNS (Dutch Lipid Clinic Network Score)	DLCNS FH diagnosztikus pontrendszer
DM (data mining)	Adatbányászat
DVT (deep vein thrombosis)	Mélyvénás trombózis
FCS (familial chylomicronemia syndrome)	Familiáris kilomikronémia szindróma
FH (familial hypercholesterolemia)	Familiáris hiperkoleszterinémia
G3PDH1	Glicerinaldehyd-3-foszfát dehidrogenáz (gén)
GOF	gain-of-function (funkció nyerő) mutáció
GPIHBP1	Glikozil-foszfatidil-inozitol kapcsolt HDL kötő protein (gén)
GPT (generative pre-trained transformer)	Transzformer alapú, neurális hálózatra épülő nyelvi modell
GPU (graphics processing unit)	Grafikus feldolgozást és megjelenítést végző számítógép chip

HC	Hiperkoleszterinémia
HIS (hospital information system)	Kórházi beteginformációs rendszer
HR (hazard ratio)	Kockázati arány (relatív hazard)
HTG	Hipertrigliceridémia
IR (information retrieval)	Információ kinyerés (adatbányászat)
LMF1	Lipoprotein lipáz érési faktor 1 (gén)
LOF	Loss-of-function (funkció vesztő) mutáció
Lp(a)	Lipoprotein (a)
LPL	Lipoprotein lipáz
MFCS (multifactorial chylomicronemia syndrome)	Multifaktoriális kilomikronémia szindróma
ML (machine learning)	Gépi tanulás
NAFLD (nonalcoholic fatty liver disease)	Nem-alkoholos zsírmájbetegség
NLP (Natural Language Processing)	Természetes nyelvfeldolgozás
PAD (peripheral arterial disease)	Perifériás artériás megbetegedés
PCSK9	Proprotein konvertáz szubtilizin/kexin type 9
PIS (patient information system)	Kórházi beteginformációs rendszer
ReLU (rectified linear unit)	Aktivációs függvény egy típusa neurális hálók esetén
ROC-AUC (area under the ROC curve)	ROC görbe alatti terület, a klasszifikációs modellek fő teljesítmény metrikája
SENS (sensitivity)	Azon pozitív esetek száma amik pozitív predikciót kaptak - klasszifikációs modellek teljesítmény metrikája
SVM (support vector machine)	Felügyelt tanulási modell (gépi tanulás egy típusa)
SZSZBMK	Szabolcs-Szatmár-Bereg megyei Kórházak
TG	Triglicerid
W2V (word2vec)	Nyelvi modellek egy típusa

Bevezetés

Az egészségügyi statisztikák alapján Európában a kardiovaszkuláris megbetegedések (cardiovascular diseases, CVD) jelentik a vezető halálokot, ugyanakkor Közép-Európában a kardiovaszkuláris okokra visszavezethető korai halálozás háromszor gyakoribb, mint Nyugat-Európában. Mindezek mellett, Magyarországon 65 éves kor után átlagosan mindössze 5 egészségben töltött év várható, míg Nyugat-Európában ugyanez az adat 12 évre tehető [1,2]. A hiperlipidémia a CVD-k egyik fő kockázati tényezője, ugyanis a magas koleszterin- és trigliceridszint hozzájárul az ateroszklerózis kialakulásához már fiatal korban is [3]. A lipidanyagcserét érintő veleszületett betegségek, például a familiáris hiperkoleszterinémia (FH) ezért jelentős kardiovaszkuláris kockázatot hordozhatnak.

Jelen dolgozatomban a lipidanyagcsere veleszületett betegségeire, azon belül is az FH és a familiáris kilomikronémia szindróma (FCS) diagnosztikai kérdéseire fókuszálok. Mindkét betegség kiemelt fontosságú abban a tekintetben is, hogy a diagnózis hiányában a szokásos gyógyszeres lipidcsökkentő kezelés hatékonysága nem megfelelő.

Az FH az egyik leggyakoribb monogénes öröklődésű betegség, melyet javarészt az low-density lipoprotein (LDL) receptor (LDLR), az apolipoprotein (Apo) B100 (Apo-B100) génjének funkcióvesztő, valamint a proprotein konvertáz szubtilizin/kexin type 9 (PCSK9) génjének funkciónyerő mutációi okoznak [4]. Korábbi adatok azt mutatják, hogy a betegség heterozigóta formája körülbelül 1:500-hoz, míg a homozigóta forma 1:1.000.000-hoz gyakorisággal alakul ki [5]. Az utóbbi évek adatai alapján betegség valós prevalenciája ettől gyakoribb, a heterozigóta forma előfordulási gyakorisága kb. 1:300, míg a homozigóta formáé 1:300 000 körülire tehető [6,7]. A világon így 20-30 millió FH beteg lehet, azonban legalább 80%-uk nem diagnosztizált. Az ismert betegeknek is csupán 10%-a részesül hatékony lipidcsökkentő kezelésben és éri el a kardiovaszkuláris rizikójának megfelelő célértékeket. Így sajnos nem meglepő, hogy a kardiovaszkuláris szövődmények következtében átlagosan 15 évvel korábban halnak meg az FH betegek, mint a génhibát nem hordozó egyének [8]. A fentieknek megfelelően, az FH jelentős kihívást jelent a kardiovaszkuláris megelőzésben hazánkban is, mivel körülbelül 20-40 ezer FH-s személy élhet Magyarországon.

Az éhomi kilomikronémia oka lehet monogénes öröklődésű ritka betegség is. Ennek során genetikai hiba okozta csökkent lipoprotein lipáz (LPL) aktivitás miatt károsodik a trigliceridekben gazdag lipoproteinek eliminációja a vérből [9]. Ez az állapot súlyosan magas

– és tartós – szérum triglicerid koncentrációhoz vezet FCS-ben. A magas triglicerid (TG) szint súlyos, potenciálisan életveszélyes állapothoz vezethet recidív hasnyálmirigygyulladással. Becslések szerint az FCS 1.000.000 emberből egyet érint és minden esetben az LPL fehérje működését jelentős mértékben károsító genetikai eltéréssel jár akár homozigóta, akár összetett (compound) heterozigóta mutációk formájában [10]. Bár az FCS betegek döntő többsége esetén homozigóta formában funkcióvesztő mutáció mutatható ki az LPL génben, szintén gyakoriak az összetett mutációk az Apo-C2, Apo-A5, lipoprotein lipáz érési faktor 1 (LMF1), glikozil-foszfatidil-inozitol kapcsolt HDL kötő protein (GPIHBP1), valamint a glicerinaldehyd-3-foszfát dehidrogenáz (G3PDH1) génekben is [11–14].

Az FCS általában már fiatal korban manifesztálódik, ellentétben a jóval gyakoribb multifaktoriális kilomikronémia szindrómával (MFCS). A magas TG szintek hátterében MFCS-ben kisebb hatású genetikai érintettség, ill. polimorfizmusok mellett egyéb, másodlagos okok is közrejátszanak, úgymint elhízás, metabolikus szindróma, 2-es típusú cukorbetegség, rendszeres alkoholfogyasztás, vagy éppen hypothyreosis és bizonyos gyógyszerek szedése. Az FCS ugyanakkor egyéb rizikófaktorok nélkül is nagyon magas szérum TG szinttel jár, ami erőteljesen fokozza az akut hasnyálmirigy-gyulladás kockázatát, melynek mortalitása 2-5% [15]. Mindezek mellett ugyanakkor az FCS betegeknél a CVD kockázata nem fokozott az átlagpopulációhoz viszonyítva. Ennek hátterében valószínűleg az LPL aktivitás jelentős csökkenése áll, amelynek köszönhetően az aterogén kilomikron és VLDL remnant partikulák kialakulása és felhalmozódása gátlődik [10]. A betegség kiemelt egészségügyi jelentősége ellenére a pontos előfordulási adatok Európa legtöbb országában nem ismertek [14].

A ritkán előforduló betegségekre, így az FH-ra és FCS-re is általánosan igaz, hogy jelentőségüket és előfordulási gyakoriságukat alábecsüljük. Diagnosztizálásuk a kórképek összetettsége és ritkasága miatt nehéz, időigényes és speciális szakorvosi tudást igényel. Fentiek alapján célkitűzésünk volt, hogy a régiókban megbecsüljük ezen betegségek pontos előfordulását és olyan hatékony diagnosztikai eszközrendszert adjunk az orvosok kezébe, amely különleges szakértelem nélkül is segítséget adhat a betegségek pontos felismerésében, a betegség hátterét és dinamikáját jelző okok és kockázati tényezők feltárásában, valamint a terápiát befolyásoló faktorok összegyűjtésében (ún. számítógéppel támogatott diagnosztika, ill. terápia). Mivel a digitalizáció fejlődésével a betegek minden releváns adata elérhető digitális és „online” formátumban, így lekérhető a kórházi információs rendszerből (HIS, hospital information system) vagy a felhőből, ezért a kutatás során főként ezeket az adatokat használtuk. A beteginformációs rendszerek (PIS, patient information system) és a HIS

számtalan jogi, informatikai és gazdasági követelménynek felelnek meg és bár a világ egyik leggyorsabban fejlődő területén, az egészségügyi informatika alkalmazása soránhasználják őket, mégis hosszú éveken keresztül állandóságot, stabilitást és kompatibilitást kell biztosítaniuk. Ezen összetett feltételrendszernek köszönhetően a világ minden országában kihívást jelent ezen komplex informatikai rendszerek struktúráltságának és flexibilitásának biztosítása, így a kutatásainkhoz elengedhetetlen volt az adatbányászat (DM, data mining) és a gépi tanulás (ML, machine learning) legmodernebb eszközeinek igénybevétele.

Irodalmi áttekintés

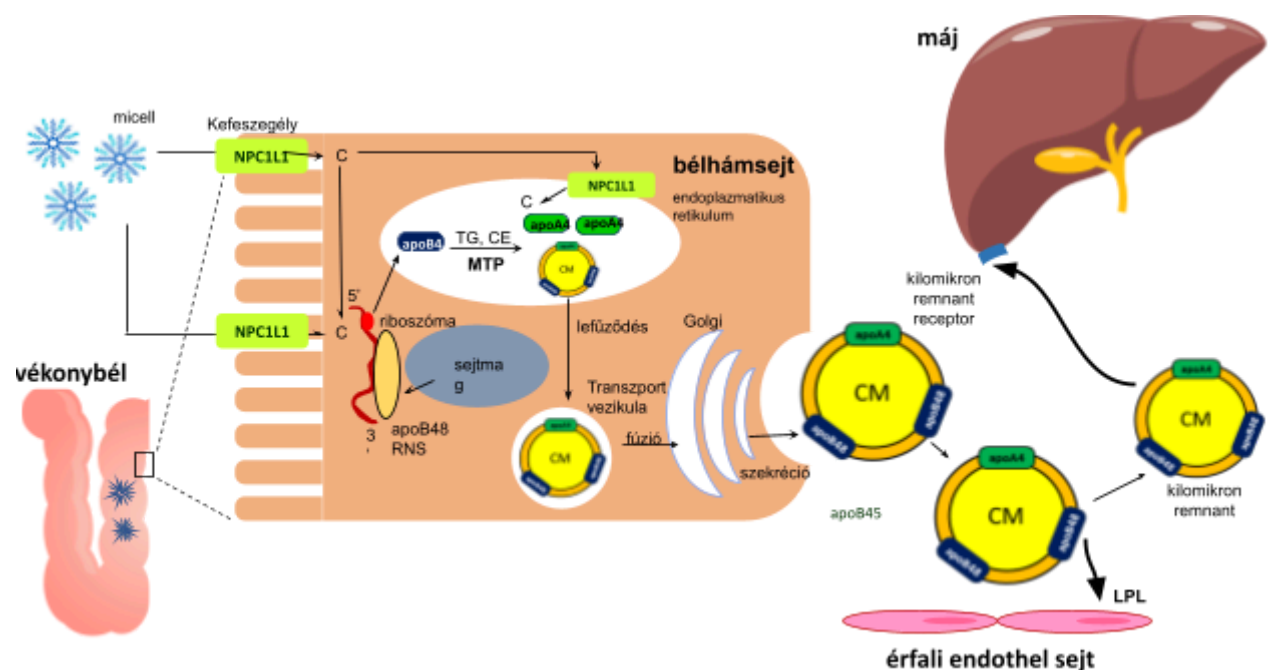
A lipidanyagcsere folyamatai alapvetően meghatározzák a szérumban a koleszterin- és trigliceridszintjét. A klinikai gyakorlatban mindkét paraméter külön-külön, vagy együttesen emelkedett volta észlelhető, amely legtöbbször a kardiovaszkuláris kockázat növekedésével, valamint az általános egészségi állapot és az életminőség romlásával jár együtt.

Exogén és endogén lipid anyagcsere

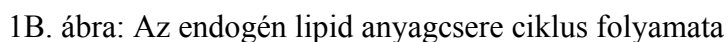
A koleszterin (chol) és triglicerid élettanilag rendkívül fontos plazmalipidek. A koleszterin funkciói rendkívül szerteágazóak, így a sejtmembrán egyik összetevője, jelátviteli fehérjék ko-lokalizációjának regulátora, valamint bizonyos hormonok és epesavak szintézisének prekursora. A lipidanyagcsereben betöltött szerepe mellett elengedhetetlenül szükséges az endokrin szabályozáshoz, vagy az idegrendszer és kritikus sejt szintű folyamatokhoz (mint amilyen az apoptózis vagy autofágia [16]), így gyakorlatilag az összes szervrendszer működéséhez. A trigliceridek pedig a szervezet fő energiaforrásaként szolgálnak, valamint a lipoproteinek másik fő lipid összetevőjeként keringenek a vérben. A főbb lipoprotein frakciók, így a kilomikron (CM), a very-low density lipoprotein (VLDL), a low density lipoprotein (LDL) és high density lipoprotein (HDL) mind fizikai, mind kémiai tulajdonságaikban, így funkciójukban is eltérnek egymástól. A lipoproteinek egymásba történő átalakulását, az általuk szállított lipidek mennyiségét és minőségét számos enzim katalizálja, melyek közül kiemelt jelentőségű a lipoprotein lipáz (LPL, ld. lejjebb), valamint a lecitin-koleszterin aciltranszferáz (LCAT) és a koleszterinészter-transzfer protein (CETP). A lipoprotein részecskék közötti koleszterinészter- és TG-cserét az utóbbi katalizálja, amelynek következtében a trigliceridben gazdag CM és VLDL, valamint a koleszterinben gazdag HDL között történik a csere. A HDL-hez és LDL-hez kötött LCAT a szabad koleszterint képes átalakítani koleszterin-észterre, mely azután a lipoproteinek magjában koncentrálódik.

A lipidanyagcsere két fő folyamatra, az ún. exogén és endogén lipidanyagcsereire bontható. Az exogén ciklus folyamata során a táplálkozással bevitt zsír emésztést követően kilomikronná alakul a vékonybél hámsejtjeiben. A kilomikronok a keringésben az érendothelhez kötött LPL által mediált hidrolízis következtében glicerint és zsírsavakat veszítenek, melyeket különböző szövetek, így pl. a zsír- és az izomszövet használ fel. A megmaradó kilomikron-remnant partikulumokat a máj veszi fel, amely a lipideket

energiatárolásra, vagy éppen epesavak és egyéb lipidszármazékok szintézisére használja fel. A májból kiinduló endogén ciklus során az ott szintetizált lipidek VLDL formájában elhagyják a hepatocitákat. A keringésben a VLDL az LPL aktivitása következtében szintén zsírsavakat veszít, ezáltal csökken a TG-tartalma és a keringésben IDL-lé, majd LDL-lé alakul. Az LDL által szállított koleszterin egyrészt a perifériás szövetek számára biztosít forrást különböző sejtfolyamatokhoz (pl. sejtmembrán felépítése, szteroidok, hormonok szintézise), emellett az ateroszklerózis kulcsmolekulája, mivel az érfalban oxidálódva és ott kumulálódva képes elindítani és fenntartani az érlemezés folyamatát. Az ateroszklerózis ellen védő mechanizmus, az ún. reverz koleszterin transzport kulcsmolekulája ezzel szemben a HDL, mely a perifériás szövetekből képes felvenni és a májba visszajuttatni a koleszterint. A fentiek ismeretében nem meglepő, hogy a CVD kialakulásának fő rizikófaktorai a magas LDL-koleszterin és az alacsony HDL-koleszterin szint. Az exogén és endogén lipidanyagcsere folyamatát az 1A. és az 1B. ábrák mutatják.



1A. ábra: Az exogén lipid anyagcsere ciklus folyamata



Etiológiai szempontból, a diszlipidémiák primer (valamilyen genetikai eltérés által meghatározott, gyakran familiáris halmozódást mutató) és szekunder (azaz valamilyen betegséghez, vagy gyógyszer-szedéshez társuló) okok köré csoportosíthatóak. A szérumban jól mérhető lipidparaméterek alapján a klinikumban igen elterjedt és egyszerű a hiperkoleszterinémia és hipertrigliceridémia, ill. a kombinált hiperlipidémia köré terjedő csoportosítás is. A vérminták elektroforézise során megfigyelhető lipoprotein-mintázat alapján pedig a diszlipidémiák öt fő csoportra oszthatóak az ún. Fredrickson-féle klasszifikáció alapján (1. táblázat) [17].

1. táblázat: A diszlipidémia Fredrickson-féle klasszifikációja és gyakoriságok

Típus	Emelkedett lipoprotein frakció	Klinikai fenotípus	Szérum chol	Szérum TG	Gyakoriság
I	Kilomikron	LPL hiány szindrómák Apolipoprotein hiány szindrómák	normál	nagyon magas	extrém ritka
IIA	LDL	Familiáris hiperkoleszterinémia Poligénus hiperkoleszterinémia	nagyon magas	normál	ritka
IIB	VLDL és LDL	Familiáris kombinált hyperlipidaemia	nagyon magas	magas	közepesen gyakori
III	IDL	Familiáris dysbetalipoproteinemia	magas	magas	nagyon ritka
IV	VLDL	Familiáris hypertriglyceridemia	normál / magas	nagyon magas	közepesen gyakori
V	Kilomikron, VLDL	Diabétész (glükóz intolerancia, hiperurikémia)	magas	magas	gyakori

Bár a klinikumban a Fredrickson tipizálás jelentősége továbbra sem elhanyagolható, mégis egyre kevésbé praktikus, mivel sem a genetikai hátteret, sem pedig a HDL szerepét nem veszi figyelembe így a klinikai gyakorlatban sokan az egyszerűbb kategorizálást részesítik előnyben [16].

Diszlipidémiák szűrése és diagnosztikája

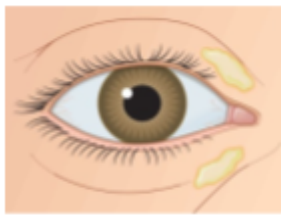
Az ateroszklerózis korai kialakulása és súlyos szövődményei miatt az FH esetében kiemelt jelentőségű a gyermekkori szűrés, hiszen a korai kezelés nélkül az érlelmeszesedés okozta kardiovaszkuláris megbetegedések (atherosclerotic cardiovascular diseases, ASCVD) kialakulása nem akadályozható meg. Ilyen esetben vagy általános szűrés történik (pl. vérvétel), vagy pedig a családi kórtörténet alapján történhet az érintett betegek felismerése, így azonban a kockázati csoportba tartozó gyermekek akár 60%-a is diagnosztizálatlan maradhat [16].

Az FCS esetében szintén fontos a gyermekkori szűrés, mert bár a betegség nem jár együtt súlyos ateroszklerózissal, mégis az akut hasnyálmirigy gyulladás rendkívül súlyos szövődményként jelentkezhet már gyermekkorban is. Fontos megemlíteni, hogy a fiatakorúak szűrésénél javasolt a gyermekkori és a fiatal felnőttkori szűrés is, mivel pubertáskorban a hormonális változások a lipidszintek gyakori változását okozhatják. A felnőttek szűrése esetén ugyanakkor a konszenzus hiányzik. Több irányelv az általános

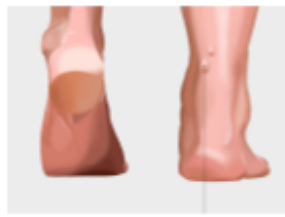
szűrést 40 éves kor alatt nem javasolja (pl. CCS - Canadian Cardiovascular Society és EAS - European Atherosclerosis Society), míg mások már 20 éves kortól javasolják az általános szűrést (pl. ACC - American College of Cardiology). Az előbbi mellett általában költséghatékonysági szempontok szólnak, míg az utóbbit a korán felismert diszlipidémia és korai intervenció támasztja alá. Fentiek tükrében egyre nagyobb jelentőséget kapnak az adathányászattal támogatott szűrések, melyek független vizsgálatok összegyűjtésére és kockázati értékek kiszámítására is alkalmasak, ezáltal az általános szűrések támogatásában, azok költséghatékonyságának növelésében is fontos szerepük lehet. A fenti betegségek kivizsgálása során a legfontosabb és leggyakoribb diagnosztikai laboratóriumi vizsgálatok a szérum lipidparaméterek – így a chol, TG, HDL-C és LDL-C – mérése. Bár az anamnézis felvétel és a fizikális vizsgálat során már a szakemberek gyanút foghatnak (pl. halmozódó és/vagy korai szívinfarktusok a családban, vagy bizonyos bőrjelenségek jelenléte, ld. alább), a legtöbb diszlipidémia klinikai tünettana mégis szegényes. Emiatt a fenti laborparaméterek rendszeres mérését javasolja a legtöbb szakmai irányelv. Az újabb előírások emellett az apoB100 szint legalább egyszeri, de inkább gyakori rutinszerű mérését javasolják [18], mivel az apoB100 keringő szintje arányos a VLDL, LDL, IDL szintekkel és az ASCVD kockázat rendkívül jó indikátora. Javasolt továbbá a lipoprotein(a) [Lp(a)] szint legalább egyszeri mérése, amely az ASCVD független kockázati tényezője.

Az alábbiakban néhány bekezdésben mutatom be az általam vizsgált súlyos, ritka primer hiperlipidémiák klinikai jellegzetességeit és diagnosztikáját.

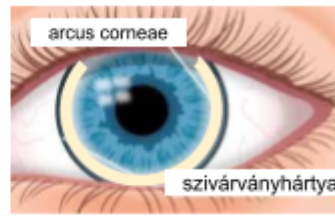
A Fredrickson IIA típusú diszlipidémiát eredményező familáris hiperkoleszterinémia igen magas koleszterin szinttel, azon belül is jelentősen emelkedett LDL-C értékkel és a kardiovaszkuláris betegségek korai kialakulásának kockázatával jellemezhető [19]. A betegség monogén öröklődésű; legtöbbször az *LDLR* vagy *APOB* gén funkcióvesztő mutációja következtében alakul ki. Előbbi következtében a májsejtek felszínén elhelyezkedő, a koleszterint a keringésből felvevő LDL-receptorok internalizációja károsodik, míg utóbbi az apoB LDL-receptorhoz való affinitását rontja – mindkét mutáció a keringésben maradó LDL-koleszterin emelkedett szintjéhez vezet. A fentiek mellett az FH lehetséges oka a *PCSK9* gén funkciónyerő mutációja lehet, melynek következtében az LDL-receptorok májsejtekben történő degradációja fokozódik [20]. A szövetekben lerakódó koleszterin miatt a betegek 20-40%-ánál megjelenik valamilyen patognomikus elváltozás is, mint pl. a tendon xanthoma az Achilles-inakon, a szemhéjak xanthelasmája, vagy a szaruhártya gyűrűs elszíneződése (arcus corneae), (2. ábra).



xanthelasma



tendon xanthoma



arcus corneae

2. ábra: Az FH külső jegyei

Az FH diagnózisában hatékony segítséget nyújt az ún. Dutch Lipid Clinic Network (DLCN) kritérium és pontrendszer [21], amely a páciensek kezeletlen LDL-C értékeit, a fizikális és a genetikai vizsgálat eredményét, valamint a saját klinikai kórtörténetüket és a családi anamnézist is figyelembe veszi. Ennek alapján a DLCN pontszám az FH diagnózisának biztos, valószínű, lehetséges és valószínűtlen voltát jelzi. A DLCN pontrendszert a 2. táblázatban mutatom be.

2. táblázat. DLCN pontrendszer az FH meghatározásához

		Pontszám
1.	Közvetlen hozzátartozó, akinek ismert korai* koszorúér- és érbetegsége van, VAGY Közvetlen hozzátartozó ismert LDL-C szintje a 95. percentilis feletti	+1
2.	Elsőfokú rokon tendon xanthomával és/vagy arcus corneae-val, VAGY 18 év alatti gyermekek, akiknek az LDL-C szintje meghaladja a 95 percentilis értéket	+2
3.	Korai* koszorúér-betegségben szenvedő beteg	+2
4.	Korai* agyi vagy perifériás érbetegségben szenvedő beteg	+1
5.	Tendon xanthoma	+6
6.	Arcus corneae 45 éves kor előtt	+4
8.	LDL koleszterinszintek a kórelőzményben (mmol/l)	
	- LDL-C ≥ 8.5	+8
	- LDL-C 6.5 - 8.4	+5
	- LDL-C 5.0 - 6.4	+3

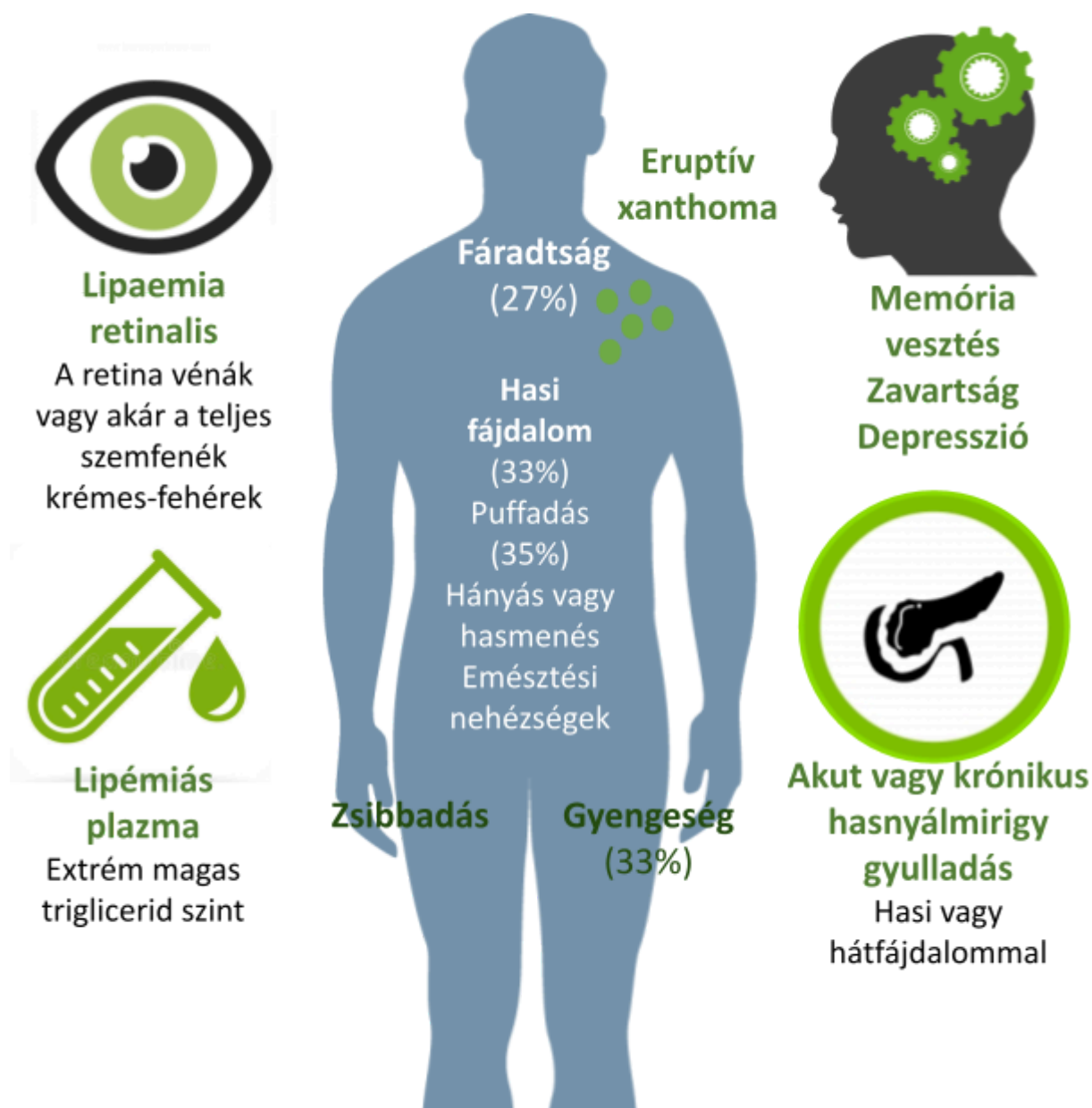
	- LDL-C 4.0 - 4.9	+1
9.	DNS-vizsgálat – funkcionális mutáció az <i>LDLR</i> , <i>apoB</i> és <i>PCSK9</i> gének valamelyikén	+1

* Korai jelentése ≤ 55 év férfiaknál; <60 év nőknél

Pontszám >8 : biztos FH diagnózis (“definite”); pontszám 6-8: valószínű FH diagnózis (“probable”); pontszám 3-5: lehetséges FH diagnózis (“possible”); pontszám <3 : valószínűtlen FH diagnózis (“unlikely”)

A dolgozatomban tanulmányozott másik ritka, monogénes eredetű hiperlipidémia a familiáris kilomikronémia szindróma, mely Fredrickson 1-es típusú diszlipidémiát (azaz jelentős hipertrigliceridémiát) eredményez; gyakoriságát mintegy 1:1.000.000-ra becsülik. Az FCS legtöbbször az *LPL* gén funkcióvesztő mutációjával, így az LPL fehérje enzimaktivitásának jelentős csökkenésével jár – így a trigliceridben gazdag lipoproteinek (a kilomikronok és a VLDL) rendkívül magas szérumszintjével karakterizálható [22]. Ahogy fentebb is részleteztem, a lipoprotein-lipáz az endogén és exogén lipidanyagcserében is alapvető szerepet játszik; működéséhez pedig számos kofaktor szükséges [22]. Az LPL enzim kulcsfontosságú aktivátoraként az apoC2 szolgál, míg a lipase maturation factor-1 (LMF1) a megfelelő fehérjeszerkezet kialakításában alapvető [23]. A kapilláris endothelhez történő szállításban és kihorgonyzásban a glycosylphosphatidylinositol-anchored high density lipoprotein-binding protein 1 (GPIHBP1) játszik fontos szerepet, az LPL-apoC2 komplexet az apoA5 stabilizálja, az apoE pedig a trigliceridben gazdag lipoproteinek eltakarítását szabályozza [24,25]. Az apoC3 és az angiopoietin-like protein (ANGPTL) 3, 4 és 8 pedig a lipoprotein lipáz gátló fehérjéiként szolgálnak [26,27]. A fenti kofaktorok együttes hatása határozza meg a lipolitikus komplex aktuális aktivitását, mely a TG degradáción keresztül a hipertrigliceridémia kialakulására jelentős hatással van.

Nem meglepő, hogy FCS-ben gyakran extrém hipertrigliceridémia figyelhető meg az exogén és endogén lipidanyagcsere károsodása és a következményesen emelkedett CM és VLDL szintek miatt. A felhalmozódó lipidek következtében az eruptív xanthoma, lipémiás plazma és a lipaemia retinalis gyakori, bár nem specifikus tünetek. A magas szérum CM szint máj- és lépmegegyesüléshoz vezethet. Az egyéb, pl. központi idegrendszeri tünetek mellett azonban az FCS legsúlyosabb potenciális szövődménye a heveny, nem ritkán ismétlődő hasnyálmirigy gyulladás, melynek mortalitása elérheti az 5-6%-ot. A súlyos hipertrigliceridémiás betegek 15%-ában legalább egyszer kialakul az akut pankreatitisz [28]. Az FCS legfőbb klinikai tüneteit a 3. ábra [22] mutatja be.



3.ábra: Az FCS leggyakoribb klinikai tünetei

A diagnosztikát támogató pontrendszer FCS esetén is rendelkezésre áll, melyet Moulin és szerzőtársai javasoltak [14]. Mivel az FCS *sine qua non*-ja a hypertrigliceridémia, így a pontrendszer (FCS score) alapja a különböző időpontokban mért TG szint. Emellett a betegek kórtörténetét (auto- és családi anamnézis), a tünetek jelentkezésének idejét (életkor) és lipidcsökkentő kezelésre adott választ is figyelembe veszi. Ezt, az FCS diagnózisára javasolt pontrendszert a 3. táblázatban mutatom be.

3. táblázat. A familiáris kilomikronémia szindróma pontrendszere Moulin és mtsai. alapján

	Pontszám
1. Éhgyomri TG > 10 mmol/L három egymást követő vérvizsgálat során	+5
Éhgyomri TG > 20 mmol/L legalább egyszer	+1
2. Korábbi TG < 2 mmol/L	-5
3. Nincs másodlagos magyarázó tényező (kivéve a terhességet és az etinilösztadiol szedését)	+2
4. Pankreatitisz az anamnézisében	+1
5. Ismeretlen, visszatérő hasi fájdalom	+1
6. Nem fordult elő családi összetett hiperlipidémia az anamnézisében	+1
7. Nem reagál a lipidcsökkentő kezelésre (TG csökkenés <20%)	+1
8. A tünetek az alábbi életkorban jelentkeznek:	
- 40 év alatt	+1
- 20 év alatt	+2
- 10 év alatt	+3

Pontszám ≥ 10 : az FCS diagnózis valószínű ("very likely"); Pontszám <10 : az FCS diagnózis valószínűtlen ("unlikely"); Pontszám <8 : az FCS diagnózis nagyon valószínűtlen ("very unlikely").

Adatbányászat

A statisztika és a különféle számítástudományok elemeit is alkalmazó adatbányászat során alapvetően kétféle feladatot különböztetünk meg. Esetünkben mindkettőre, így a prediktív és a deskriptív feladatokra is szükség volt a kutatásaink során. A prediktív feladatok lényege egy függő változó ("target", vagy "dependent" variable) értékének meghatározása. Jelen esetben ez főként egy bináris klasszifikációs feladat, azaz a betegség meglétének vagy hiányának megállapítása, amely számos független változóból ("independent", vagy "explanatory" variable) áll [29]. Ezen változóknak a meghatározása, transzformálása, normalizálása, stb. szintén a feladat része (ld. lentebb). Ezzel szemben, a deskriptív feladatok célja az adatban rejlő mélyebb összefüggések feltárása, megértése, vizualizálása, így a

korrelációk, klaszterek és adatkapcsolatok feltárása, melyet rendszeresen használtunk a kutatásaink során felhasznált adatok és összefüggések generálásához.

Szöveges adatok feldolgozása

A szövegbányászat (text mining), mely az adatbányászat egy külön ága, mely a szöveges adatok feldolgozásával foglalkozik. A szövegbányászat interdiszciplináris területként magában foglalja a számítástudomány, a statisztika, valamint a nyelvészet tudományos eredményeit és technikáit. Célja a strukturálatlan adatok feldolgozása és az információ hatékony kinyerése különféle technikák alkalmazásával. Az információkinyerés (IR, information retrieval), mint folyamat, meghatározott kifejezésekkel, algoritmusokkal és keresésekkel dolgozik (queries), mely keresések adathoz vagy információhoz történő hozzáférést kezdeményeznek adattáblákból, adatbázisokból vagy egyéb strukturálatlan, ill. strukturált adathalmazból. Mindezek alapján a nyers, rendezetlen szöveghez strukturált információt rendel. Az IR fontos alapfolyamata a "tokenizáció" (tokenization), a szűrés (filtering) és a "stemming" (szótövezés), melyek a rendezetlen szövegek előfeldolgozásának fontos lépései [30]. Tokenizáció során a szöveget kisebb alkotóelemekre, így mondatokra és szavakra bontjuk, szűrés során a szövegből bizonyos elemeket, karaktereket (pl. írásjeleket, speciális karaktereket, ún. stop szavakat) távolítunk el, míg a szótövezés során a szavakat megszabadítjuk a hozzájuk kapcsolódó elő- és utótagoktól [31]. A szövegbányászat fontos eleme még az ún. "regexek", azaz reguláris kifejezések (regular expressions) alkalmazása. A regex olyan szintaktika, mely lehetőséget ad szabályok és mintázatok egyszerű leírására, így alkalmazható a már előfeldolgozott strukturálatlan szövegeken, egyszerűbb és összetettebb kifejezések keresésére is. A regex alkalmazásával igen bonyolult kifejezések logikája is előállítható, valamint a helyesírási hibák és elírások problémája is kezelhető, illetve a legtöbb programnyelven és programozói környezetben rendkívül gyorsan futattható: ezért a regexek alkalmazása a szövegbányászat egyik legalapvetőbb eszköze [32].

Mindezek mellett a dokumentumok, illetve kifejezések rangsorolása is egyike a legfontosabb szövegbányászati feladatoknak. Ennek egyik legtöbbször használt technikája, a TF-IDF (Term Frequency - Inverse Document Frequency), amely a dokumentumokban használt szavak, kifejezések előfordulási gyakoriságát vizsgálva következtet azok fontosságára, relevanciájára. Ezek alapján minden kifejezés-dokumentum pároshoz egy értéket rendel, amely alapján a dokumentumok összehasonlíthatóvá válnak (emellett ezek az értékek a keresés problémája során is jól hasznosíthatók). Egy kifejezés TF-IDF értéke

arányosan nő a kifejezés előfordulási gyakoriságával a kategorizált dokumentumokban, ugyanakkor az érték kompenzálódik azoknak a dokumentumoknak a számával, ahol ez a kifejezés előfordul (így például a szövegben gyakran használt kötőszavak alacsony TF-IDF értéket kapnak) [31].

A gépi tanulás során a szöveges adatok feldolgozásának legfejlettebb és legösszetettebb módja a természetes nyelvfeldolgozás (NLP, natural language processing) gépi tanulás alapú módszerei, melyek az írott vagy beszélt nyelvi struktúrának a reprodukálását célozzák és ezáltal alkalmasak szövegek összetett morfológiai, szintaktikai (mondattani) és szemantikai (jelentéstani) elemzésére. Az NLP segítségével így az egyes kifejezések a gép számára a kontextuson belül értelmezhetővé válnak. A nyelvi struktúra reprodukálása nyelvi modelleken (ún. sztochasztikus, azaz véletlenszerű modelleken) keresztül történik, melyek nyelvi struktúrák létrehozására képesek a szövegeken tanult szabályszerűségek figyelembevételével. Az utóbbi időkben leginkább neurális hálózatokat használnak erre, de léteznek statisztikai, szabály-alapú megközelítések is, melyek elsősorban a korábbi, 80-as, 90-es évek nyelvi feldolgozására voltak jellemzőek. A statisztikai nyelvi megközelítések közül a leghíresebb az n -gram nyelvi modell (bag-of-words, unigram, bigram, trigram, stb.), ami a szöveget és annak alkotóelemeit (például a mondatokat) a szavak vagy szócsoporthoz együttes előfordulási valószínűségén keresztül írja le [33]. Az unigram modellek nem tesznek különbséget a szavak között (a következő szó előfordulási valószínűsége független az öt megelőző szó, vagy szavak valószínűségétől), míg a magasabb rendű modellek figyelembe veszik az adott szó kisebb-nagyobb kiterjedésű környezetét (a következő szó előfordulási valószínűsége az öt megelőző szó, vagy szavak együttes előfordulási valószínűségéből számolandó). Nem meglepő, hogy minél nagyobb környezetet veszünk figyelembe, annál nagyobb számítási kapacitást igényel az ilyen nyelvi modell felállítása.

A becsült és a tényleges előfordulási valószínűség közötti eltérést perplexitásnak nevezzük, ami így a nyelvi modellek egyfajta teljesítménymetrikájának is felfogható. Azaz, minél alacsonyabb a perplexitás, annál sikeresebb egy nyelvi modell az adott szöveg struktúrájának megtanulásában [33]. A modern neurális háló alapú nyelvi modellek a szöveget sokdimenziós vektorterek formájában írják le. A word2vec modellek esetén minden szó egy-egy vektorként értelmezhető, a szavak leírása a vektortérben pedig egy sekély, kétrétegű neurális háló segítségével történik (a folyamatot szóillesztésnek, vagy “word embedding”-nek nevezzük). A vektorok közötti távolság (belső szorzat, “cosine similarity”) a word2vec modellek esetén felfogható a szavak közötti szemantikai hasonlóság mértékének is

[34]. A word2vec modelleket a transzformer alapú modellek haladták túl a 2010-es évek közepén, melyek a neurális hálókra épülő nyelvi modellek közül jelenleg a legsikeresebbek. A transzformer alapú modellek két fő előnye, hogy figyelembe veszik a szavak mondaton belüli pozícióját és a többi szóhoz való kapcsolatát is (“position encoding” és “self-attention”); valamint többrétegű struktúrájukból adódóan (többrétegű, de azonos súlyú encoder-decoder struktúrák) minden szó esetén parallel számítást tesznek lehetővé, így különösen hatékonyan taníthatóak sokszálú (GPU - Graphics Processing Unit) és sokchipes elosztott cloud rendszereken [35]. A legjelentősebb transzformer alapú nyelvi modellek a BERT (Bidirectional Encoder Representations from Transformers), mely egy “encoder-only” modell, amit 2018-ban publikáltak a Google fejlesztői (és azóta több mint 150 publikáció foglalkozott a modell továbbfejlesztésével és elemzésével) [36], valamint a GPT (Generative Pretrained Transformer) [35]. A GPT alapvetően struktúrában a BERT-hez hasonló, de azzal ellentétben generatív, azaz képes a nagy mennyiségű címkézetlen (“unlabeld”) szövegen tanulva saját szövegek előállítására.

Bináris klasszifikáció

A bináris klasszifikáció (osztályozás) során a tárgyat, azaz a bemeneti mintát (objektumot) matematikai műveletek segítségével két kategória egyikébe soroljuk (pl. 0: nem beteg; 1: beteg). A folyamat során tulajdonságok egyfajta halmazához ($\mathbf{x}$) egy klasszifikációs modell segítségével címkét rendelünk (y). A klasszifikációs tanulási feladat inputja egy $(\mathbf{x}, y)$ rendezett n -es véges lista (ún. “tuple”), a klasszifikációs feladat pedig legtöbbször leíró (deskriptív) vagy előrejelző (prediktív) [29]. A klasszifikáció matematikailag-statisztikailag tulajdonképpen a regressziószámítás speciális esetének is tekinthető (ahol a “függő” változó mindig diszkrét értékeket vehet csak fel). Jelen fejezetben a bináris klasszifikáció ismertetése során követem a Tan, Pang-Ning által szerkesztett alapvető adatbányászati ismereteket bemutató szakkönyv “Introduction to Data Mining - (Classification: Basic Concepts and Techniques)” harmadik fejezetének a logikáját [29].

A klasszifikációs technika vagy algoritmus egy olyan rendszerszerű megközelítés, mely során az input adatokból, tanuló algoritmusok segítségével egy osztályozási (klasszifikációs) modellt építünk, amely a lehető legpontosabban képes leírni a $(\mathbf{x})$ és (y) közötti kapcsolatot (tréning halmaz, vagy “training set”), valamint szintén képes helyesen kategorizálni olyan adatokat, amelyek nem voltak részei a modellépítésnek (teszt halmaz, vagy “test set”). A klasszifikáció pontossága és így a matematikai modell sikeressége ennek

megfelelően egy ún. igazságmátrix (“confusion matrix”) segítségével, valamint az igazságmátrixból (4. táblázat) levezethető teljesítménymetrikákkal (5. táblázat) értékelhető (a legtöbb tanuló algoritmus esetén a cél vagy a pontosság maximalizálása, vagy pedig az ezzel logikailag ellentétes hibaarány minimalizálása).

4. táblázat: A bináris klasszifikáció igazságmátrixa

		Predikció (osztályok)	
		1	0
Tényleges érték (osztályok)	1	f_{11} = tényleges pozitív (TP)	f_{10} = fals negatív (FN)
	0	f_{01} = fals pozitív (FP)	f_{00} = tényleges negatív (TN)

5. táblázat: A bináris klasszifikáció leggyakrabban használt teljesítménymetrikái

Metrika	Számítás	Szöveges összefoglaló
Pontosság (ACC)	$\frac{TP + TN}{\text{Összes eset}}$	A helyes predikciók aránya az összes predikcióhoz viszonyítva
Precizitás (PPV)	$\frac{TP}{TP + FP}$	Az összes pozitívnak jósolt eset közül a ténylegesen pozitívak aránya (pozitív prediktív érték)
Szenzitivitás (TPR)	$\frac{TP}{TP + FN}$	A ténylegesen pozitív esetek hány százaléka kapott pozitív predikciót. Egyéb elnevezés: Hit rate, Recall
Specificitás (TNR)	$\frac{TN}{TN + FP}$	A ténylegesen negatív esetek hány százaléka kapott negatív predikciót. Egyéb elnevezés: Szelektivitás
Diagnosztikai hatékonyság (DOR)	$\frac{TP \cdot TN}{FP \cdot FN}$	Annak az esélye, hogy egy pozitív eset ténylegesen pozitív kategóriát kap a negatív predikcióhoz képest
Hibaarány (ER)	$\frac{FP + FN}{\text{Összes eset}}$	A hibás jóslatok aránya az összes predikcióhoz viszonyítva

A bináris klasszifikáció talán legfontosabb és legismertebb összetett teljesítménymetrikája a ROC-AUC görbe [29]. A ROC (Receiver Operating Characteristics) egy valószínűségi görbe, amely különféle küszöbértékek mellett mutatja a bináris szeparáció hatékonyságát, míg az AUC (Area Under The Curve) összességében mutatja a szétválasztás sikerességét. A két mértékre sokszor összességében, mint AUROC (Area Under the Receiver Operating Characteristics) hivatkoznak. A ROC görbe tulajdonképpen a 0 és 1 közötti FPR (false positive rate - a ténylegesen negatív esetek hány százaléka kapott pozitív predikciót)

értékekhez (x tengely = 1-specificitás) rendel 0 és 1 közötti TPR (y tengely = szenzitivitás) értékeket, míg az AUC ezen görbe 0 és 1 közötti integrálja. A modellenként eltérő AUROC magasabb értéke a pozitív és negatív esetek közötti szeparáció sikerességét mutatja: 0,5 esetén ez a szeparáció véletlenszerű, míg 1 esetén egyértelmű. Amennyiben az integrál értéke 0,5 alatti, akkor pedig egy fordított döntéssel kapunk sikeres modellt, aminek a sikeressége az 1-AUC értékkel ekvivalens.

A klasszifikációs modellek ("classifier"-ek), azaz az osztályozó algoritmusok széles skálája érhető el (pl. döntési fák, neurális hálózatok, support vector machine, stb.), azonban ezek részletes ismertetése pedig jócskán túlmutat jelen dolgozat keretein. A következőkben a modellek főbb csoportjait mutatom be néhány szóban ismertetve őket.

A **döntési fák** (klasszifikációs vagy regressziós fák) az egyszerűbb osztályozó modellek közé tartoznak. Exponenciális skálázódásuk miatt az optimális fák általában nem meghatározhatóak, így közelítő algoritmusok szükségesek az építésükhöz (a legtöbb technika a Hunt-féle rekurzív algoritmusra épül, mely a nagyobb entrópiájú attribútumok, azaz a felismerést könnyebben lehetővé tulajdonságok felől halad a kisebb entrópiájú, tisztább attribútumok irányába a fa építése során. Az entrópia meghatározása a folyamat során mérőszámok segítségével történik (pl. Gini-index). A döntési fák csomópontokból ("node") és ágakból állnak. A csomópontok közül kiemelt jelentőségű a végpont, ahol a klasszifikáció történik ("leaf nodes" = levelek), valamint a kezdőpont, ahonnan a fa indul ("root node", vagy "root" = gyökér). A döntési fák legnagyobb problémája a túltanulás, melynek kezelésére többféle módszer is kidolgozásra került (pl. "pre-pruning", "post-pruning").

Az **NN klasszifikáció** ("k-nearest neighbors" classifier) a legjelentősebb késleltetett tanulóalgoritmus ("lazy learner"). A "lazy learner"¹ (szemben az eddig bemutatott, azonnali "eager learner" módszerekkel szemben) azt jelenti, hogy a betanítási adatok általánosítása nem történik meg a klasszifikáció előtt, ugyanakkor fontos megemlíteni, hogy a "lazy learner" nem jelenti azt, hogy a számításigény kisebb lenne, csupán itt a számításigény a klasszifikáció közben történik, nem pedig előtte az általánosítás során ("eager learner" algoritmusok esetében a modell alkalmazása előtt történik egy matematikai általánosítás - függvényillesztés, mely eredménye alapján történik a klasszifikáció - a "tanulóadatok" a klasszifikáció során közvetlenül nem kerülnek felhasználásra). Ez különösen a nagyméretű és gyakran változó adathalmazok esetén hasznos. Az NN klasszifikáció során megnézzük, hogy egy d-dimenziós állapotterben egy vizsgált x pont környezetében található k darab eset

¹ Itt a "tanulás" megnevezés félrevezető, hiszen itt nem történik igazi tanulás. A modellparaméterek nem változnak és nincs általánosítás sem.

milyen kategóriákba tartozik. A leggyakoribb (esetenként távolsággal súlyozott) kategória lesz a vizsgált eset kategóriája [37,38].

A **Bayes-klasszifikáció** (Bayes classifier) a feltételes és kapcsolt valószínűségek viszonyát leíró Bayes-tételre épül. A tanuló adaton meghatározott “*a priori*” független valószínűségek segítségével a Bayes-tétel alapján kiszámítható, hogy egy kategorizálatlan eset milyen valószínűséggel esik egy adott kategóriába. A Bayes-klasszifikáció nagyon hatékony minden olyan adaton, ahol a kezdeti valószínűségeloszlás jól meghatározható és a multikollinearitás alacsony (azaz az egyes modellbeli változók közötti kapcsolat nem jelentős, azaz tényleg egymástól függetlennek tekinthetők); emellett szintén jól alkalmazható a hiányzó értékekre és az irreleváns attribútumokra. A független valószínűségi változók meglepte sokszor egy nagyon erős megkötés, hiszen a klasszifikációk nagy részénél kisebb-nagyobb korreláció megfigyelhető. Ezekben az esetekben jól használhatóak a Bayes-hálózatok (BBN, Bayesian Belief Networks), ahol a függőségi kapcsolatok mértéke és azok iránya is hatékonyan jeleníthető meg egy gráf segítségével és a hálózat interpretálása is egyszerű. Hátránya ezeknek a rendszereknek, hogy - amennyiben nem ismertek - a helyes gráfok megtalálása nehéz feladat [39].

A **mesterséges neurális hálózatok** (ANN, artificial neural networks) a gépi tanulás legnépszerűbb és leggyorsabban fejlődő modellei. Egy ANN csomópontokból (“neuron”) és a csomópontokat összekötő kapcsolatokból áll (hasonlóan az idegrendszer neuron-axon-dendrit kapcsolataihoz). A legegyszerűbb ANN a perceptron, ami több input neuron és egy output neuron rendszere. A neuronok aktiválása aktiváló függvények segítségével történik, melyek meghatározzák az ANN típusát: ilyen aktiváló függvény pl. a Softplus, Sigmoid vagy a ReLU (legegyszerűbb esetekben szignumfüggvény). A legtöbb ANN többrétegű azaz tartalmaz egy vagy több rejtett réteget (“hidden layers”) egy vagy több rejtett neuronnal (“hidden nodes”). A több rejtett réteggel rendelkező neurális hálókat mély neurális hálóknak is nevezzük (innen ered az elnevezés “deep learning”), míg az egyetlen rejtett réteggel rendelkező modelleket sekély (“shallow learning”) neurális hálóknak. Az ANN lehet előrecsatolt (“feed-forward”), amennyiben a neuronok csak a következő réteg neuronjaival állnak kapcsolatban vagy rekurrens (“recurrent”), amennyiben egymással, vagy több réteg neuronjaival is kapcsoltak. Az ANN tanítása során leggyakrabban a hiba-visszaterjesztés (“backpropagation”) módszerét alkalmazzuk, a függvényparamétereket pedig a gradiens módszer (“gradient descent”) elve szerint optimalizáljuk. Az ANN általános approximátor, így gyakorlatilag bármilyen valós függvény becslésére alkalmas, viszont különösen érzékeny a bemeneti zajra és a túltanulásra. A gradiens módszer sokszor a globális minimum helyett

lokális minimumokat talál, mely több matematikai módszerrel kezelhető (egzakt megoldása jelenleg nem ismert a problémának), de ez legtöbbször tovább növeli az ANN amúgy is jelentős tanítási idejét különösen sok rejtett réteg és neuron esetén.

Az **SVM** (support vector machine, vagy support vector network), mely sekély neurális hálónak is tekinthető, a sekély tanulási módszerek talán legnépszerűbbike. Az SVM az SVC (support vector classifier, vagy soft margin classifier) továbbfejlesztése azáltal, hogy egy magfüggvény ("kernel function") segítségével a tanulási halmaz pontjai között magasabb dimenziójú kapcsolatokat is figyelembe vesz és úgy határozza meg az adatokat elválasztó hipersíkot, azaz egy magasabb dimenziójú SVC-t. Az SVM nagy előnye, hogy a tanulási probléma konvex optimalizálási probléma, szemben az ANN, vagy a fák heurisztikus megközelítésével, így a megtalált minimum egyben biztosan globális minimuma is a célfüggvénynek.

Az **ensemble**, vagy kombinált **klasszifikációs** megoldások során azt használjuk ki, hogy több, egymástól részben független osztályozó algoritmus esetén az együttesen figyelembe vett döntés során a klasszifikációs hiba valószínűsége kisebb, mintha csak egyetlen algoritmust használnánk. Az ensemble módszer lényege, hogy több modellt építünk az eredeti adathalmazból és a döntéseiket aggregáljuk. Ezt alapvetően négyféleképpen tehetjük: a tréning halmaz alakításával ("bagging" és "boosting" módszerek), a bemeneti tulajdonságok változtatásával (melyek különösen sok felesleges változó esetén sikeresek, például a "random forest" módszerek), a címkézés módosításával (akkor hatékony, amennyiben egy esethez nagyon sok címke tartozik), illetve a tanuló algoritmusok változtatásával (ami különösen sikeres a heurisztikus, nem stabil modellek alkalmazásakor - például fák, neurális hálók esetén).

Célkitűzések

A ritka anyagcsere betegségek tanulmányozása során az alábbi feladatok megoldására törekedtünk:

- az FH betegek regionális prevalenciájának meghatározása adatbányászattal;
- az FCS betegek regionális prevalenciájának meghatározása adatbányászattal;
- az FH betegek kardiovaszkuláris kockázati tényezőinek felmérése;
- az FH betegek szérum Lp(a) értékeinek vizsgálata és összehasonlítása a nem FH populációval, adatbányászati módszerekkel;
- a FH betegek, különösen a “valószínű FH” csoport pontosabb azonosítása gépi tanulási módszerekkel;
- az FCS betegek gyors és automatizált szűrése, az FCS pontszám gyors becslése adatbányászati és gépi tanulási módszerek alkalmazásával;
- az FCS betegek egyéb rutin laborvizsgálatok alapján történő, hatékony azonosítási lehetőségének vizsgálata adatbányászat és gépi tanulási módszerek alkalmazásával;
- annak megállapítása, hogy a gépi tanulás és adatbányászat eszköztára mennyire képes a mindennapi orvosi munka és a klinikai kutatások támogatására.

Anyag és módszer

A kutatásaink során a vizsgált betegségek megértéséhez, prevalenciájuk becsléséhez és diagnosztika támogatásához a hagyományos adatbányászat (DM, data mining) eszköztárát használtuk fel. Az adatok feldolgozása során az észak-alföldi régió két, nagy forgalmú egészségügyi intézményéből származó, a klinikai és kórházi adatbázisokból kinyert, anonimizált adatot használtunk. A Debreceni Egyetem Klinikai Központban és a Nyíregyházi Jósa András Egyetemi Oktató Kórházban 2007. január 1. és 2014. december 31. között megjelent, közel másfél millió ($n=1.342.124$) beteg összes, a fenti két intézményben keletkezett orvosi dokumentumából származó, tisztított és strukturált, anonimizált adatbázist használtunk. A nyolc teljes évet átölelő adatforrásként a laboratóriumi eredmények és a kórházi/klinikai megjelenések során keletkező szöveges adatok (ambuláns lapok és zárójelentések) mellett a betegségek nemzetközi osztályozását biztosító International Statistical Classification of Diseases and Related Health Problems- (ICD)-10-re transzformált diagnosztikai adatokat használtunk.

Az eredmények értékelése és feldolgozása a tudományos munkánk során az alábbiak szerint történt:

1. FH és FCS betegek **regionális prevalenciájának** a becslése adatbányászati módszerekkel a már ismert klinikai diagnosztikai kritériumok alapján.
2. FH **betegek** Lp(a) értékeinek, valamint metabolikus paramétereinek, kórelőzményének és szövődményeinek **összetett vizsgálata** távoli adattáblák és adatok közötti gyors kapcsolatok biztosításán és ezáltal leíró statisztikák gyors generálásán keresztül.
3. FCS betegek **klinikai diagnózisa** számítógéppel támogatott előszűrés alapján.
4. Javaslatok megfogalmazása az FCS pontrendszer hatékonyabbá tételéhez, ezáltal az **FCS hatékonyabb diagnosztikájához mesterséges intelligencia segítségével.**

A betegek szűrésének kiinduló lépése a klinikai adatok egységesítése, megfelelő formátumba történő alakítása és tisztítása volt, hogy azokon már hatékonyan tudjuk futtatni az informatikai szkripteket és algoritmusokat. Mivel a kutatáshoz több egészségügyi szolgáltató több évnnyi adata is felhasználásra került, így ezek a lépések különösen fontosak voltak egyrészt az adatok mennyisége miatt, mivel a hibás és inkonzisztens adatrészek

lényegesen rontják a rajtuk futó algoritmusok teljesítményét és futási idejét. Másrészt pedig az egy személyhez tartozó több különféle adatforrás közötti konzisztencia megteremtése, illetve azok összefűzése és egységesítése nélkül a feldolgozási hiba is nagyon magas lenne, amely különösen ritka betegségek esetén (amikor is a páciensek számossága alacsony) a kiértékelési módszertanok használhatóságát is megakadályozhatja.

Alapvetően 5 típusú klinikai adattal kerültünk kapcsolatba a kutatásaink során, melyek az alábbiak voltak:

- A betegvizsgálat szöveges adatai: anamnézis (kórelőzmény), epikrízis (rövid kórtörténeti kivonat), műtéti leírás (az elvégzett beavatkozások rövid leírása);
- A laborvizsgálatok numerikus adatai és a vizsgálatok egyéb fontos paraméterei (például a referencia tartományok, mértékegységek, a vizsgálatokat végző gépek adatai);
- A feltételezett és igazolt diagnózisok adatai;
- A beavatkozások és vizsgálatok eredményei;
- A betegek egészségügyi rendszertől független statisztikai adatai (pl. születési hely és idő, nem, végzettség, foglalkozás).

A klinikai adatforrásokon belül, egy betegre vetítve természetesen sok duplikáció és inkonzisztencia található: egyrészt azért, mert az adat sokféle csoportosításban, sok egészségügyi szereplő számára elérhető, akik szintén rögzítenek a már meglévő adathoz, illetve módosításokat hajtanak végre azokon; másrészt azért, mert a betegút során a különféle beavatkozási pontokon más-más információkkal és intézményi prioritásokkal rendelkeznek, illetve eltérő feltételezésekkel élnek az egészségügyi szereplők. Alapvetően ezek az adat természetesen velemjárói és egyszerre tudják nehezíteni és könnyíteni az adatfeldolgozást, de emiatt a feldolgozás során kezelni kell őket. Így például a duplikált diagnózisok gyakoriak és hasznosak, mert könnyebb egy diagnózis megtalálása és egy páciens kategorizálása, de az inkonzisztens diagnózisok megléte – pl. egyszerre 1-es és 2-es típusú diabetes mellitus diagnózisa – nehezíti a tisztánlátást és sokszor részletesebb elemzéseket és mélyebb feltárásokat igényel.

Ahogy fent is részleteztem, az egészségügyi adatbázishoz egy előfeldolgozási folyamat utáni nyers szöveges formában fértünk hozzá, amely már átesett egy anonimizálási fázison, így az adatok és a tényleges páciensek személye már nem volt összekapcsolható. A

további tisztítási és előfeldolgozási folyamatok részben nyílt forrású eszközök segítségével (<http://pandas.pydata.org/>, <http://www.numpy.org/>), részben saját fejlesztésű Python 2.8.X és Python 3.X szkriptek segítségével történtek (<https://www.python.org/>). Végeredményként a hibás adatelemek javításra és feltöltésre kerültek, valamint az adatok egységes és homogén formában tárolódtak: a szöveges adatok “utf-8” kódolású szöveggént, a számok számként, míg a dátumok UTS vagy python dátum formátumként. Így nagy, összefüggő és indexelt adatfájlokat kaptunk, melyek minden orvosi eset diagnózisát, laboratóriumi vizsgálatát, a beavatkozások és diagnosztikai eljárások leírásait, valamint az összes szöveges adatot tartalmazta. Az egyének személyével összefüggő anonim statisztikai adatai külön táblában kerültek tárolásra. Az előfeldolgozási folyamat során különféle szerializálási és pufferezési módszereket is alkalmazni kellett a masszív adattömeggel kapcsolatos memória problémák kiküszöbölésére. Ezen feladatokat elsősorban a python alapú pandas, a unix környezetben népszerű awk (<https://www.gnu.org/software/gawk/>) illetve a dask (<https://www.dask.org/>) framework-ök segítségével oldottuk meg. Az előfeldolgozási folyamatok végén az adat PostgreSQL 13.x adatbázis táblákba került rögzítésre és a továbbiak során innen került lekérdezésre.

A kórházi adatoknál több probléma is előfordul, így az adattisztítás esetünkben sem volt elkerülhető. Az adatok összegyűjtése és elemzése során megállapítottuk, hogy az általunk feldolgozott adatokban az irodalomban előforduló, az adatminőséget befolyásoló főbb tényezők legtöbbje megtalálható volt:

6. táblázat: Főbb adathibák a vizsgált egészségügyi adatokban

Adathibák típusai	Példa
Zaj és műtermékek	Főleg laborvizsgálatok és képi diagnosztika
Kilógó értékek (“outlierek”)	Laborvizsgálatok, mért értékek (pl. BMI)
Hiányzó értékek	Anamnézis, szöveges adatok
Inkonzisztens értékek	Alapvető páciens statisztikák (pl. születési év)
Duplikátumok	Korábbi esetek megjelenése új eseteknél (főleg szöveges adatok), kórházak közötti adatok

Az adathibák kezelhetőek voltak a feldolgozás során megfelelő módszerek alkalmazásával, így legtöbbször akár figyelmen kívül hagyva őket, vagy megbecsülve a hibás adatok helyes értékét más, akár duplikált adatforrások alkalmazásával, vagy hibajavító mechanizmusokkal. Szintén fontos, hogy az adattal és annak vizsgálatával eltöltött hosszú idő során komoly adattudás halmozható fel, ami elengedhetetlen a leggyakoribb hibák gyors kijavításához. Szintén fontos kérdés az adat tárgyalásakor, hogy mennyire használható az egészségügyi ellátás során keletkező adat, melyet időbelisége és relevanciája is jellemez. Az ilyen adatok, információk nagy része hosszú ideig tartó rögzítés és kutatások eredményeképp jött létre, így szerencsére lassan avulnak el. Azonban így is előfordulnak olyan esetek, ahol idővel az adatok egy része már nem alkalmazható (például a gyógyszeres kezelések korszerűsödése következtében), illetve előfordulhat, hogy korábbi adatokon bizonyos műveletek csak korlátozottan használhatóak (például egy új genetikai diagnosztika esetén). Relevancia tekintetében (azaz, hogy az elérhető adat tartalmaz-e elegendő információt az adott problémára történő alkalmazáshoz) a legfontosabb megemlíteni a “sampling bias” jelenségét, így azt a kérdést, hogy az egészségügyi intézményekben megjelenő populáció vajon mennyire reprezentálja jól a teljes lakosságot. Ennek becslése kiemelt jelentőséggel bírt a kutatásaink során.

Adattisztítás adatbányászattal

A szöveges adatok strukturálásához és a bennük szereplő információ kinyeréséhez további feldolgozási lépések szükségesek (“preprocessing”). A fontos kifejezések megtalálásához és csoportosításához parsing (a parszolás során egy szöveges adatot elemzünk, hogy előre meghatározott struktúrára hozzuk, vagy előre meghatározott struktúra szerint értelmezzük), stemming, stop word szűrés, BOW (bag-of-words, részletesen az Irodalmi Áttekintés fejezetben) modellezés és kifejezések rangsorolása (TF-IDF, W2V - Keras keretrendszerben, <https://keras.io/>) történt [40,41]. A BOW modellek a szöveg kifejezéseit, azok előfordulási gyakoriságát hisztogramként adják meg a szekvenciális struktúra (azaz a szövegen belüli sorrend) figyelembe vétele nélkül. A W2V modellek emellett már a szekvenciális struktúrát és a kontextust is figyelembe veszik, az egyes kifejezések egy vektortérben jelennek meg. Mindkét modell alkalmas arra, hogy kifejezések fontosságát természetes módon meghatározzuk velük, az IDF érték és perplexitás alapján. A vizsgálataink során emellett orvosok segítségével is meghatároztunk egyes, a betegségekre jellemző kifejezéseket, melyeket ezután regexek segítségével kerestünk a betegek

kórtörténetében. A fenti módszerekkel a kezdeti többmilliós kifejezéstér két-százötvenezerre csökkentettük, valamint sikerült a legfontosabb kifejezések (például dohányzás vagy káros alkoholfogyasztás) meghatározása és azok diszkretizálása is. Ezek a lépések elengedhetetlenek voltak az egyes pontrendszerek (DLNCS, FCS score) anamnézist és családi kórtörténetet érintő elemeinek pontos meghatározásához. Ugyanakkor az előfeldolgozás itt még nem ért véget, mivel ahhoz, hogy hatékonyabban megtaláljuk a szövegben elrejtett elemeket és szöveg alapján hatékonyan tudjuk az eseteket összehasonlítani további nyelvi feldolgozásra, ismert nyelvi modellekre és saját magunk által tanított nyelvi modellekre volt szükségünk. A nyelvi feldolgozáshoz az NLTK (Natural Language Toolkit, 3.x verzió) Python könyvtárat használtuk. Az alapvető szöveges előfeldolgozás során elérhetőek voltak a legfontosabb szóbeágyazódások (“word embedding”), unigramok, bigramok és n-gramok. Mivel ezeket a beágyazódásokat magyar nyelven csak klasszikus szövegtészleten (korpusz) tanították, így mindenképp saját nyelvi modelleket kellett tanítanunk. Ehhez egy ún. GRU (Gated Recurrent Unit) alapú modellt fejlesztettünk [42], emellett az általános kifejezések esetében a BERT (Bidirectional Encoder Representations from Transformers) modellt használtuk. Ez alapján a végső és kritikus, legfontosabb kifejezések kétféleképpen kapcsolódhattak egy orvosi esethez: egyrészt, az elsődleges feldolgozás, vagy regex alapján egyértelműen azonosítható volt az esethez egy kifejezés, másrészt pedig az alkalmazott nyelvi modellek segítségével az adott orvosi esetnél a kifejezés magas valószínűséggel volt megtalálható.

Az adattisztítás végeztével a klinikai adatok halmazát (“dataset”) egy adatmátrixba rendeztük. Az adatmátrix páciens (eset) szinten tartalmazta a legfontosabb attribútumokat, valamint a lényeges attribútumokhoz tartozó folytonos (pl. laborértékek), diszkrét (pl. hipertónia: igen/nem) és sorrendi (pl. diagnózisok sorrendje) értékeket.

Összefoglaló a statisztikai elemzésekről

A statisztikai elemzések Python által támogatott adatbányászati és statisztikai szoftvercsomagokkal történtek, döntően Conda 4.X környezetben, munkánk során Python 3.8, IPython 7.2x, Cython 0.2x, Pandas 0.2x és Numpy 1.2x szoftvereket használtunk. Elsősorban a prediktív modellek tanítása, illetve kiértékelése során szükség volt egy “betegségmentes” (az adott diagnózissal biztosan nem rendelkező) betegpopuláció létrehozásához, melybe olyan egyének tartoztak, akik az alábbi betegségek, kórállapotok egyikével sem rendelkeztek: elhízás, dohányzás, cukorbetegség, aktív lipidcsökkentő (sztatin)

kezelés, ill. a szérumban mért magas LDL-C és alacsony HDL-C szintek. Utóbbiak esetén az adott egészségügyi intézményben elfogadott határértékeket meghaladó szinteket tekintettük magasnak, vagy alacsonynak. A szignifikancia számítása minden esetben 95%-os szignifikancia szinten és nem párosított t-próbával történt. A statisztikai ábrákat Matplotlib 3.5-ös szoftvercsomaggal és kapcsolódó könyvtárakkal generáltuk (Seaborn).

Páciensek szűrése FH és FCS diagnózisra

A DLCN score (DLCNS, vagy DLCN érték) meghatározása előtt kiszámoltuk a betegek átlagos lipidértékeit [43], minden olyan esetet kiszűrve, ahol a lipidcsökkentő kezelés már ismert volt a fent részletezett nyelvi feldolgozás alapján. Ezt követően, a hozzá tartozó orvosi esetek összessége alapján, a már tisztított és előfeldolgozott adatokon minden egyes betegre vonatkozóan kiszámoltuk a DLCN értéket, ezáltal megbecsülve az FH valószínűségét. A betegek így az (i) az FH diagnózisa biztos, (ii) az FH diagnózisa valószínű és (iii) az FH diagnózisa lehetséges besorolást kaptak.

A tisztított és előfeldolgozott adatokon történt az FCS score-ok kiszámítása is. Az így kiszámított és az FCS valószínűsége alapján csoportosított adat szakorvosi értékelés alá került, majd a klinikai diagnózis után gépi tanulási módszerekkel (ReLU neurális hálózat, adaptív boosting, gradient boosting, SVM) került további feldolgozásra és tréningre. A modellek tanításához nyílt forráskódú szoftvereket használtunk. Az AdaBoost, SVM implementáció során scikit-learn 0.23 (<https://scikitlearn.org>), míg a ReLU hálókhoz PyTorch 1.6 (<https://pytorch.org>), a fákhhoz pedig XGBoost 1.2.1 (<https://xgboost.readthedocs.io/en/latest/>) programokat alkalmaztunk. A táblázatokban a paraméter keresés során talált legjobb eredményeket jelenítették meg. A kísérletek során véletlenszerűen két egyenlő részre, 50-50%-ban tanuló- és tesztalmazra bontottuk az adatot.

Gépi tanulás

Az FCS esetek azonosítása egy bináris klasszifikációs problémának tekinthető, mivel egy véges, valós számértéket felvevő leíró vektor (laborvizsgálatok, kórtörténet, diagnosztika, stb.) és egy bináris címke (0: az illető nem FCS beteg; 1: az illető FCS beteg) érhető el minden vizsgált egyén esetén. Többféle módszer is létezik arra, hogy kapcsolatot teremtsünk az állapottér és a címkék között. Ahhoz, hogy ezeket a módszereket megvizsgáljuk és kiválasszuk a legjobban teljesítőt, részmintákat képeztünk a teljes populációból ismert és diagnosztizált FCS betegek adatai alapján, majd az ezeken feltanított

matematikai modelleket független, de szintén ismert és diagnosztizált FCS betegekkel rendelkező adathalmazokon értékeltük ismert teljesítmény metrikák alapján (ROC-AUC, ACC, SENS). A fenti érvelés alkalmazhatóságát a Vapnik-Chervonenkis féle generalizációs elmélet adja [44], mely a mélytanulás kivételével minden bemutatott modellre használható.

Mélytanulás esetén az uniform konvergencia, mely az alapja a Vapnik-Chervonenkis tételeknek, nem feltétlenül magyarázza a gyakorlatban tapasztalt generalizációt, lásd pl. Nagarajan és Kolter [45], de más kisebb komplexitású modell esetében alkalmazható többek között a tanulóhalmaz méretének kiválasztásához, lásd Devroye és munkatársai [46]. Az elmélet alapján három követendő szabály létezik a sikeres modellek kiválasztásához: (i) alacsony komplexitású modellek preferálása [47], (ii) független teszhalmazon történő értékelés és (iii) a lehető legnagyobb tanuló-, vagy tréninghalmaz alkalmazása. Mindezzel összhangban, hogy a különböző, de legsikeresebb modelleket teszteljük az FCS, mint ritka és kis elemszámú betegség azonosításához, három – széles körben alkalmazható – matematikai modellt használtunk, melyek az ensemble fák (tree ensembles) [47,48], a “sekély” kernel alapú (SVM) [49] és “mély” ReLU aktivációs neurális hálók [50]. Utóbbi hálózatok megfelelően nagy szélesség és mélység esetén jó approximátorok bármilyen valós, folytonos függvény esetében az adatok, megfigyelések alapján a háttérben húzódó mintázat megtalálásában. Különösen jól teljesítenek spatiotemporális adatokon, valamint ha ismert ismétlődő struktúrák vannak az adatban és ha a tulajdonságokat tartalmazó adatmátrix nem tetszőleges sorrendiségű. Mivel az FCS diagnózisok esetén a struktúrára semmilyen speciális feltétel nem teljesül, így teljesen kapcsolt (“fully connected”, “struktúra-agnosztikus”) neurális hálókkal dolgoztunk, ReLU aktivációs függvényvel és adaptív momentum (Adam) optimalizációval [51].

Az FCS diagnózisok kis adathalmazok, melyeknél az adatmátrix tulajdonságok sorrendisége nem meghatározó, így már elsőre is valószínűsíthető, hogy az ensemble fák és a kernel-alapú modellek hatékonyabban fognak működni az osztályozásban. Az ensemble fák gyenge osztályozók (“weak classifier”-ek) halmazai, melyek kisméretű, csaknem véletlen döntési fákból állnak, itt a tanítás során adaptív boosting [47] (ADA) és gradient boosting [48] (XGB) módszereket alkalmaztunk. A kernel alapú módszerek közül a legismertebbet, az SVM-et alkalmaztuk lineáris, polinomiális és radiális bázisfüggvény alapú kernelekkel. A tanuló- és teszhalmazok mért eredményeket modellekre teljesítményeit a 7-es táblázat mutatja be.

7. táblázat: Klasszifikációs teljesítmény metrikák az FCS betegeken tanított modellek esetén

Tréning halmaz	Teszt halmaz	Módszer	Kísérlet	átlagos AUC	szórás AUC	átlagos ACC	szórás ACC	átlagos SENS	szórás SENS	átlagos SPEC	szórás SPEC
50%	50%	ReLU	30	0.735	0.064	0.895	0.024	0.212	0.160	0.950	0.029
Diagnosztizált esetek (1) és magas FCS score (n=275) (0)		SVM	30	0.792	0.054	0.927	0.013	0.0	0.0	0.999	0.001
		ADA	30	0.770	0.053	0.902	0.014	0.110	0.121	0.970	0.023
		XGB	30	0.810	0.042	0.909	0.018	0.070	0.104	0.976	0.025
		ReLU	30	0.599	0.088	0.857	0.112	0.237	0.184	0.859	0.113
Teszteléshez random 5000 klinikai eset hozzáadása FCS diagnózis nélkül.		SVM	30	0.872	0.057	0.998	0.001	0.0	0.0	0.999	0.001
		ADA	30	0.824	0.092	0.996	0.002	0.110	0.121	0.999	0.002
		XGB	30	0.871	0.074	0.997	0.001	0.070	0.104	0.999	0.001
		ReLU	30	0.906	0.041	0.997	0.001	0.245	0.142	0.999	0.011
Tréning halmazhoz 1000 random klinikai eset (0) hozzáadása FCS diagnózis nélkül. Teszteléshez random 5000 klinikai eset hozzáadása FCS diagnózis nélkül.		SVM	30	0.955	0.024	0.999	0.001	0.0	0.0	0.999	0.001
		ADA	30	0.923	0.051	0.996	0.002	0.110	0.121	0.999	0.001
		XGB	30	0.982	0.015	0.997	0.001	0.091	0.096	0.999	0.001
		ReLU	30	0.906	0.041	0.997	0.001	0.245	0.142	0.999	0.011

A legfontosabb teljesítmény metrikák bináris osztályozásához a szenzitivitást (azaz, a tényleges FCS betegek hány százalékát mondja a modell FCS betegnek), a specificitást (azaz, a nem FCS betegek hány százalékát mondja a modell nem FCS betegnek) és a pontosságot (azaz, azon betegek arányát a teljes sokaságban, akik pontos osztályozást kaptak függetlenül attól, hogy betegek-e vagy sem) választottuk a ROC-AUC görbe mellett. A ROC görbe definíció szerint a TPR (szenzitivitás) és FPR (1-specificitás) pontpárokkal felrajzolt görbe az összes valószínűségi küszöbérték (0-1) mellett ("probability threshold"). Az AUC érték (mint a ROC görbe integrálja) úgy interpretálható, mint annak a valószínűsége, hogy egy pozitív mintát előrébb sorolunk konfidencia alapján, mint egy negatív mintát. Megemlítendő, hogy az adatmátrix tulajdonságainak fontossága többféleképpen meghatározható a tanított modellek alapján, míg SVM és mélytanulás esetén ezen bemeneti változók szerepének becslése nem egyértelmű. Gradient boosting esetén ezek a tulajdonságok a fákon betöltött pozíciójukból egzaktan is meghatározhatók. Számítalan módszer biztosított a fontossági sorrend megtalálásához a megjelenések egyszerű megszámlálásától a komplex megoldásokig. A kutatás során a fontossági sorrend megállapításához elsősorban az előző módszer (boosting) súlyozott változatát használtuk [52].

Laborvizsgálatok

A laboratóriumi vizsgálatok mindig az adott egészségügyi intézmény aktuális protokolljának megfelelően történtek, nem feltétlenül éhgyomorra. Ugyanakkor, az Lp(a) szintek tizenkét órás éhezést követően, éhgyomorra levett vérérszérumból kerültek meghatározásra a Debreceni Egyetem Klinikai Központ Laboratóriumi Medicina Intézetében Cobas c501 Analizátor (Roche Ltd., Mannheim, Németország) segítségével a gyártói előírásoknak és protokollnak megfelelően. Az Lp(a) analízis referencia tartománya <300 mg/L volt.

Eredmények

Kutatásaink során vizsgáltuk az FH és FCS betegek előfordulási gyakoriságát, az FCS betegek hatékonyabb diagnosztikájának lehetőségeit. Részletesen elemeztük az FH betegek patogenezisét, különös tekintettel az Lp(a) szintek változására és a kórlefolyásra kifejtett hatására. Eredményeinket három közleményben publikáltuk.

Regionális FH és FCS prevalencia

Az FH és az FCS diagnózisát elősegítő pontrendszereket, így a DLCN és az FCS score-okat kétféleképpen is felhasználtuk a prevalenciák becsléséhez. Egyrészt a rendszerek mögötti orvosi logika figyelembevételével a diagnózisok, tünetek és laborértékek alapján alcsoportokat képeztünk, másrészt pedig minden egyes beteg esetén, a rá vonatkozó összes elérhető adat alapján kiszámoltuk az adott betegség valószínűségét jellemző pontszámot. Minderre azért is volt szükség, mert bár a diagnosztikai problémák klasszikus bináris problémák, addig a prevalenciák klinikai adattömegeken történő becslése már statisztikai folyamatnak tekinthető, ahol az adott egyénekről csak korlátozott információ áll rendelkezésre. A kétféle megközelítésből adódó különbségek további hasznos információkat is szolgáltathatnak, mint pl. az összetartozó csoportok egyikéhez való tartozás erős valószínűséget szolgáltat a másik csoporthoz való tartozáshoz is abban az esetben is, ha az adott adat az adott személyhez nem elérhető.

A familiáris hiperkoleszterinémiában szenvedő betegek azonosításához, így az FH prevalencia becsléséhez, a kardiovaszkuláris rizikócsoportok figyelembe vételével kilenc betegcsoportot képeztünk, melyeket az életkoruk és a lipidparamétereik alapján is jellemeztünk (8A. és 8B. táblázatok).

8A. táblázat: A betegek megoszlása kardiovaszkuláris rizikófaktorok és életkor alapján

Csoportosítás		Életkor (év)				Összes egyén (n)
		+18 (n)	19-30 (n)	♀31-60 ♂31-55 (n)	♀61+ ♂56+ (n)	
1.csoport	FH az anamnézisben	11	10	69	45	135
2.csoport	Hiperkoleszterinémia az anamnézisben	80	357	13 428	13 552	27 387
3.csoport	CAD az anamnézisben	67	363	10 245	16 883	27 558

4.csoport	CVD az anamnézisben	167	608	15 779	33 416	49 970
5.csoport	PAD az anamnézisben	23	109	13 373	48 450	61 955
6.csoport	Tendon xanthoma az anamnézisben	1	2	5	7	15
7.csoport	Arcus corneae az anamnézisben	3	8	35	34	80
8.csoport	LDL-C > 3,4 mmol/L és TG < 1,7 mmol/L	71	282	6 563	4 559	11 475
9.csoport	CHOL >5,2 mmol/L és TG < 1,7 mmol/L	107	492	9 441	6 918	16 958

8B. táblázat: A betegek megoszlása kardiovaszkuláris rizikófaktorok és lipidparaméterek alapján

Csoportosítás		Lipid paraméterek		
		Összkoleszterin (mmol/l)	LDL-C (mmol/l)	HDL-C (mmol/l)
1.csoport	FH az anamnézisben	7,90 ± 1,66	5,56 ± 1,55	1,41 ± 0,29
2.csoport	Hiperkoleszterinémia az anamnézisben	6,69 ± 1,61	4,71 ± 1,37	1,44 ± 0,44
3.csoport	CAD az anamnézisben	5,47 ± 1,55	3,48 ± 1,29	1,26 ± 0,40
4.csoport	CVD az anamnézisben	6,27 ± 1,61	4,70 ± 1,28	1,42 ± 0,43
5.csoport	PAD az anamnézisben	6,39 ± 1,82	4,91 ± 1,26	1,45 ± 0,43
6.csoport	Tendon xanthoma az anamnézisben	5,26 ± 0,90	3,15 ± 0,63	1,47 ± 0,52
7.csoport	Arcus corneae az anamnézisben	5,69 ± 1,75	3,81 ± 1,54	1,45 ± 0,25
8.csoport	LDL-C > 3,4 mmol/L és TG < 1,7 mmol/L	7,67 ± 1,01	5,45 ± 0,77	1,60 ± 0,42
9.csoport	CHOL >5,2 mmol/L és TG < 1,7 mmol/L	7,90 ± 1,02	5,47 ± ,91	,168 ± 0,46

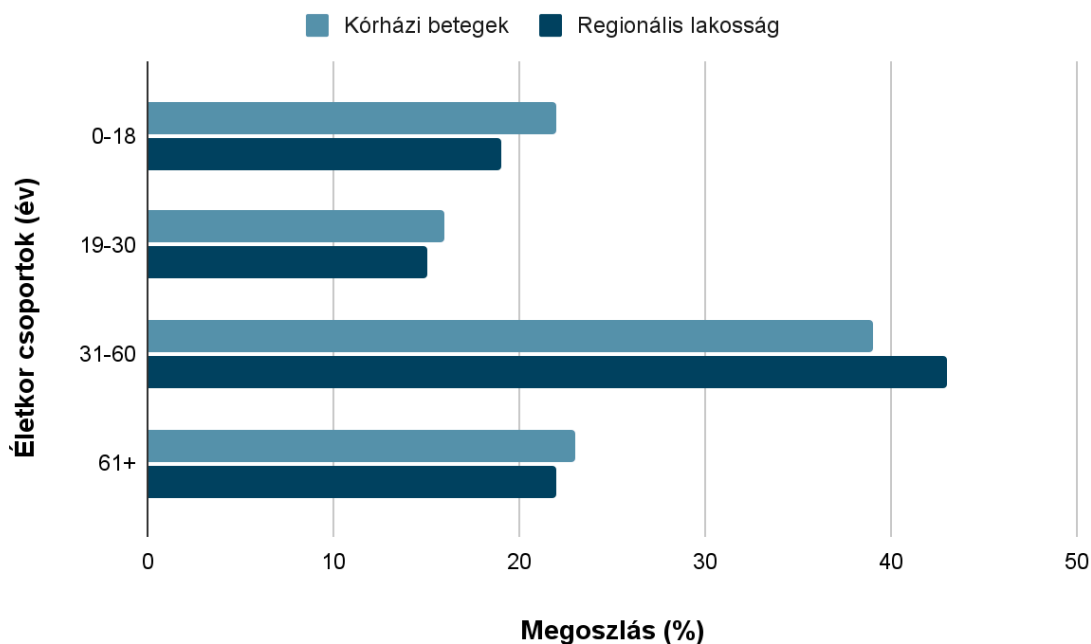
Az első csoportba kerültek azok, akiknél biztos FH diagnózis volt felállítható a szöveges adatbányászat segítségével, míg a másodikba kerültek azok, akiknek a kórtörténetében magas koleszterinszint szerepelt. A harmadik, negyedik és ötödik csoport azon egyéneket tartalmazza, akiknél fellelhető volt a CAD (coronary artery disease,

koszorúér-betegség), CVD (cerebrovascular disease, agyi érrendszeri megbetegedés), valamint a PAD (peripheral arterial disease, perifériás artériás betegség) diagnózisa. A hatodik csoportba a tendon xanthoma diagnózisával, a hetedikbe pedig az arcus corneae diagnózisával rendelkező betegek kerültek. A nyolcadik csoportba a magas LDL-C- és alacsony trigliceridszinttel rendelkező páciensek, míg a kilencedik csoportba a magas összkoleszterin- és alacsony trigliceridszinttel jellemzett betegek kerültek. Minden egyes betegre kiszámoltuk a DLNC pontértéket és ez alapján soroltuk őket biztos, valószínű és lehetséges FH kategóriákba (9. táblázat).

9. táblázat: A Dutch Lipid Clinical Network Score kritériumai szerint kategorizált betegek lipidparaméterei

	FH biztos	FH valószínű	FH lehetséges	Összes (n)
DLCN érték	>8	6-8	3-5	≥ 3
Betegek száma (n)	6	225	11 706	11 937
Arány (%)	0,001	0,017	0,87	0,89
CHOL (mmol/l)	$10,3 \pm 0,7$	$9,38 \pm 3,0$	$7,34 \pm 1,2$	-
LDL-C (mmol/l)	$8,1 \pm 0,6$	$7,61 \pm 2,4$	$5,26 \pm 0,8$	-
HDL-C (mmol/l)	$2,42 \pm 0,4$	$1,54 \pm 0,6$	$1,58 \pm 0,4$	-

Eredményeink alapján a régió betegeinek 0,89%-át érintheti a familiáris hiperkoleszterinémia. Ugyanakkor kérdéses, hogy az egészségügyi intézményekben bármilyen okból megjelenő populáció mennyire képezi le a régió teljes populációját. Mivel Magyarországon a teljes lakosság közel 76%-a átlagosan évente egyszer megfordul az állami egészségügyben és mivel a vizsgált populáció epidemiológiai adatai lényegében egyeznek a 2011-es regionális népszámlálási adatokkal (4. ábra), a klinikai prevalenciát mérvadónak tekinthetjük a tényleges prevalencia becslése során. Korábbi tanulmányban a 3 vagy a feletti DLSN értékkel rendelkező pácienseknek felajánlott genetikai vizsgálat 33%-ban erősítette meg genetikailag az FH diagnózist [53] Mindezek alapján, számításaink szerint minden 340 emberből 1 érintett hazánkban FH-val.



4. ábra: Egészségügyi intézményben megjelenők és a regionális népesség demográfiai adatai az észak-alföldi régióban

Az értékelésünk nem vette figyelembe a DLNCS genetikai és családi anamnézisre vonatkozó elemeit (ez később az FCS esetén is felmerül problémaként), mivel ezek az adatok csak rendkívül korlátozottan voltak elérhetőek a klinikai és kórházi adatbázisban. Ugyanakkor, ezt a hibát részben ellensúlyozta az, hogy régiókban és hazánkban relatíve magas a lakosság közegészségügyben való éves megjelenése, emellett pedig gépi tanulási módszerekkel próbáltuk a fenti hiányosságokat ellensúlyozni. Így a különböző DLNCS értékkel rendelkező egyének adatai a gépi tanulás során inputként szolgálnak és segítségükkel vizsgálható, hogy az egyes pontszámok esetén milyen páciens összeállítással lehet a legjobb tanulást, azaz a legmagasabb ROC-AUC görbét elérni – mindezeket többféle gépi tanulási modellel is tesztelve (mély ReLU hálók, SVM, boosting tree). Megemlítendő, hogy inputként az összes fontos, klinikailag elérhető adatot, így a laborvizsgálatok numerikus adatait, az ICD-10 diagnózisokat, valamint az NLP segítségével előfeldolgozott szöveges adatokat is felhasználtuk.

A legjobb tanulás esetén feltételeztük, hogy a csoport a lehető legjobban különül el a többiektől: így valószínűleg a csoporton belüli betegek egységes tulajdonságokkal rendelkeznek, mely feltételezhetően az FH. Mindezek alapján megállapítható volt, hogy amennyiben egy vizsgált egyén FH pontszáma 7-8, vagy afeletti, akkor szinte bizonyosan FH

beteg. Így összesen 459 beteg volt azonosítható, mely alapján arra következtethetünk, hogy a régióban minden 1300-adik ember (459:590.500) biztosan familiáris hiperkoleszterinémiában szenved.

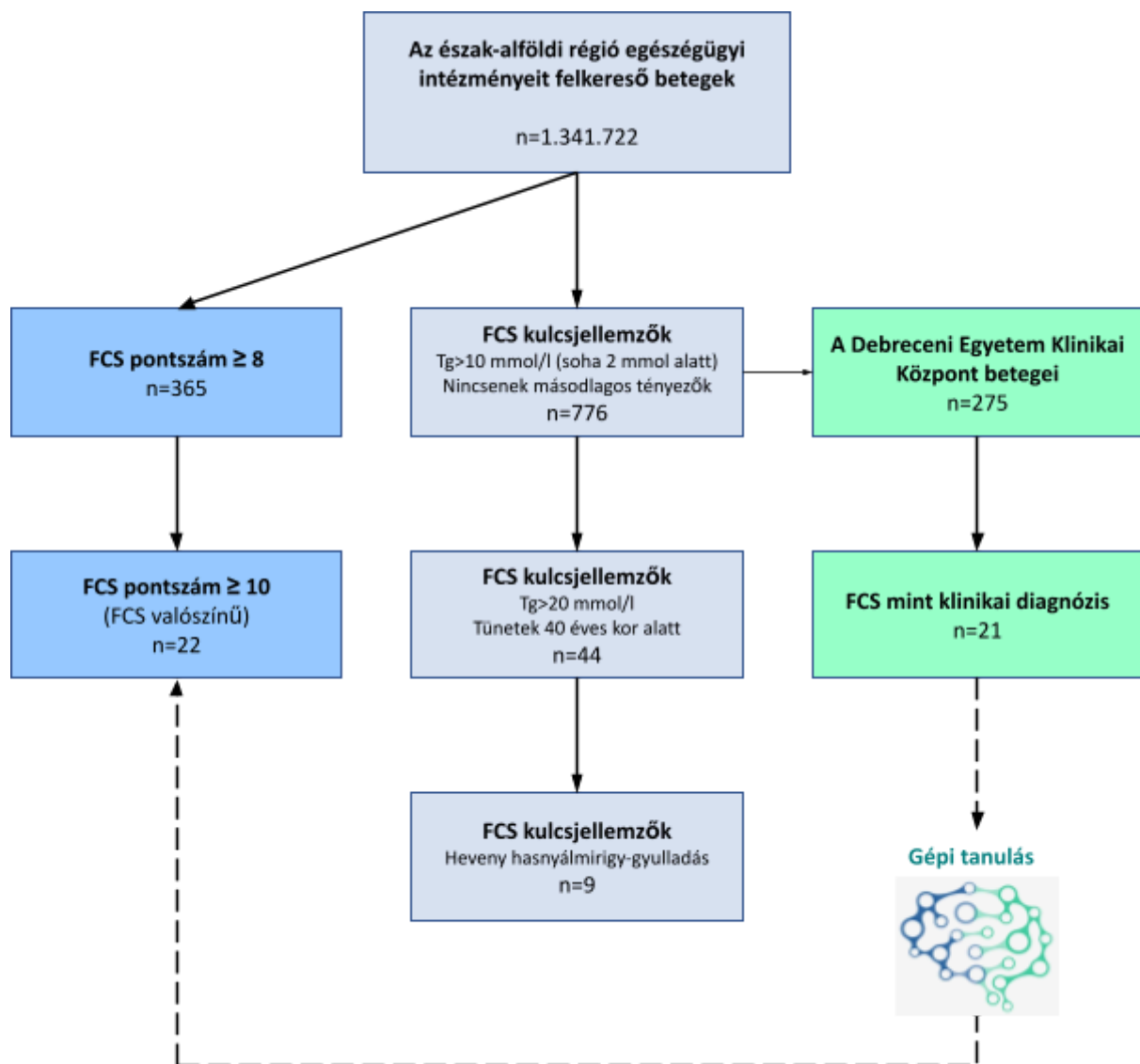
Az általunk vizsgált másik ritka kórkép, azaz a familiáris kilomikronémia szindróma esetén is minden betegre kiszámoltuk az FCS pontszámot. Ahogy fentebb ismertettük, a Moulin és mtsai. által javasolt, jelenleg elfogadott kritériumrendszer alapján kiszámolt, ≥ 10 pont esetén minden egyént FCS betegnek tekintettünk. Így összesen 26 beteg volt azonosítható, mely alapján az FCS regionális prevalenciája 19,4:1.000.000-ra becsülhető (10. táblázat).

10. táblázat: Az észak-alföldi egészségügyi szolgáltatókat felkereső betegek FCS pontszámai (pcm = 1:100.000; ppm = 1:1.000.000)

FCS klaszter	FCS pontszám	Férfi betegek	Női betegek	Összes beteg	Betegek százalékos megoszlása
Nagyon valószínűtlen	0+	602.258 (45%)	739.464 (55%)	1.341.722	100%
	1+	5.612 (56%)	4.334 (44%)	9.946	7,41‰
	2+	1.659 (75%)	558 (25%)	2.217	1,65‰
	3+	1.441 (75%)	493 (25%)	1.934	1,44‰
	4+	1.307 (74%)	461 (26%)	1.768	1,32‰
	5+	1.272 (74%)	453 (26%)	1.725	1,29‰
	6+	909 (78%)	254 (22%)	1.163	8,67‰
	7+	705 (79%)	182 (21%)	887	6,61‰
Valószínűtlen	8+	298 (82%)	67 (18%)	365	2,72‰
	9+	56 (81%)	13 (19%)	69	5,14 pcm
Nagyon valószínű	10+	17 (77%)	5 (23%)	22	1,64 pcm
	11+	3 (75%)	1 (25%)	4	2,98 ppm

Fontos megemlíteni, hogy az ilyen nagy esetszámú adathalmazok sok szempontból nehezen kezelhetőek, így az FCS prevalencia gyors becsléséhez olyan, ún. kulcsjellemzőket próbáltunk meghatározni, melyek a nagy esetszámot gyorsan nagyságrendekkel alacsonyabb esetszámmra csökkentik (“vágják”). Ritka anyagcsere betegségeknél gyakori egy-egy laborérték extrém eltérése, amely lehetőséget biztosít az egyének gyors, nagymértékű, nyers szűrésére. Kiindulásképpen a már fentebb bemutatott, Moulin és mtsai. által javasolt FCS score rendszert és a betegség mögött meghúzódó orvosi logikát alkalmaztuk, mely

tudásanyagot adatbányászati és gépi tanulási módszerekkel integráltuk. A folyamat összefoglalása az 5. ábra bal oldali paneljein látható.



5. ábra. Az FCS pontrendszer gyors becslésének folyamata

A gyors becsléshez magas FCS pontszámú betegeket választottunk ki. A betegség kulcsjellemzőiként a pontszámba nagymértékben, azaz nagy pontszámmal beleszámító, valamint jól mérhető és nem szubjektív elemeket használtunk. Ilyen kulcsjellemző volt a (i) 10 mmol/l-t meghaladó szérum TG-szint három egymást követő vérvizsgálat során (5 pont), mely (ii) soha nem volt dokumentálhatóan 2 mmol/l alatt (mely 5 pont levonást jelentett volna); (iii) nincs jelen másodlagos tényező (2 pont); a (iv) legalább egy alkalommal 20 mmol/l felettinek mért TG-érték; a (v) 40 éves kor előtt jelentkező tünetek; valamint a (vi) hasnyálmirigy-gyulladásra vonatkozó pozitív kórtörténet.

Mindezek alapján a Debrecenben megjelenő betegek között az FCS prevalenciája 8,47 ppm-nek, míg a nyíregyházi intézményben 5,32 ppm-nek adódott (11. táblázat).

11. táblázat: Az FCS pontszám becslése kulcsjellemzők alapján

A) Debrecen, minden beteg*				
FCS valószínűség	Kulcsjellemzők	FCS pontszám	Betegek száma	Betegek aránya
Nagyon valószínűtlen	Összes beteg	0+	590.500	100%
	TG 10+ mmol/l és soha 2- mmol/l	5+	665	1,13‰
	Nincsenek másodlagos tényezők **	7+	275	4,67‰
Valószínűtlen	TG 20+ mmol/l legalább egyszer	8+	85	1,44‰
	Tünetek jelentkezése 40 éves kor előtt	9+	24	4,06 pcm
Nagyon valószínű	Akut hasnyálmirigy-gyulladás	10+	5	8,47 ppm
B) Nyíregyháza, minden beteg*				
FCS valószínűség	Kulcsjellemzők	FCS pontszám	Betegek száma	Betegek aránya
Nagyon valószínűtlen	Összes beteg	0+	751.624	100%
	TG 10+ mmol/l és soha 2- mmol/l	5+	1.046	1,39‰
	Nincsenek másodlagos tényezők **	7+	501	6,67‰
Valószínűtlen	TG 20+ mmol/l legalább egyszer	8+	93	1,23‰
	Tünetek jelentkezése 40 éves kor előtt	9+	20	2,66 pcm
Nagyon valószínű	Akut hasnyálmirigy-gyulladás	10+	4	5,32 ppm

(A): * Azok a betegek, akik 2007–2014 között legalább egyszer megjelentek a Debreceni Egyetem Klinikai Központjában; ** cukorbetegség, metabolikus szindróma, pajzsmirigy alulműködés, kortikoszteroid terápia, alkoholfogyasztás.

(B) * Azon betegek, akik 2007–2014 között legalább egyszer megjelentek a Szabolcs-Szatmár-Bereg Vármegyei Oktatókórházban; ** cukorbetegség, metabolikus szindróma, pajzsmirigy alulműködés, kortikoszteroid terápia, alkoholfogyasztás.

Az egyes egészségügyi szolgáltatókat külön-külön vizsgálva, az FCS pontszámot az adatbázisban hozzáférhető minden egyes betegre is kiszámoltuk (12. táblázat). Mindezek alapján a 10, vagy afeletti FCS pontszámmal (azaz az FCS nagyon valószínű diagnózisával) rendelkező betegek előfordulási gyakorisága Debrecenben 27,11 ppm-nek, Nyíregyházán pedig 13,3 ppm-nek adódott.

12. táblázat: A vizsgált egészségügyi intézményekben megjelenő betegek számított FCS pontértékeinek megoszlása

FCS valószínűség	FCS pont	Férfiak (n)	Nők (n)	Összes beteg (n)	Betegek aránya
A) Debrecen, minden beteg*					
Nagyon valószínűtlen	0+	251.949 (43%)	338.149 (57%)	590.098	100%
	1+	2368 (53%)	2.108 (47%)	4.476	7.59‰
	2+	589 (74%)	208 (26%)	797	1.35‰
	3+	538 (73%)	198 (27%)	736	1.25‰
	4+	506 (73%)	188 (27%)	694	1.18‰
	5+	490 (73%)	183 (27%)	673	1.14‰
	6+	340 (76%)	107 (24%)	447	7.58‰
	7+	250 (78%)	71 (22%)	321	5.44‰
Valószínűtlen	8+	110 (77%)	32 (23%)	142	2.41‰
	9+	31 (82%)	7 (18%)	38	6.44 pcm
Nagyon valószínű	10+	10 (77%)	3 (23%)	13	2.20 pcm
	11+	2 (67%)	1 (33%)	3	5.08 ppm
B) Nyíregyháza, minden beteg*					
Nagyon valószínűtlen	0+	251.949 (43%)	338.149 (57%)	590.098	100%
	1+	2368 (53%)	2.108 (47%)	4.476	7.59‰
	2+	589 (74%)	208 (26%)	797	1.35‰
	3+	538 (73%)	198 (27%)	736	1.25‰
	4+	506 (73%)	188 (27%)	694	1.18‰
	5+	490 (73%)	183 (27%)	673	1.14‰
	6+	340 (76%)	107 (24%)	447	7.58‰
	7+	250 (78%)	71 (22%)	321	5.44‰
Valószínűtlen	8+	110 (77%)	32 (23%)	142	2.41‰
	9+	31 (82%)	7 (18%)	38	6.44 pcm
Nagyon valószínű	10+	10 (77%)	3 (23%)	13	2.20 pcm
	11+	2 (67%)	1 (33%)	3	5.08 ppm

(A): * Azok a betegek, akik 2007–2014 között legalább egyszer megjelentek a Debreceni Egyetem Klinikai Központjában

(B) * Azon betegek, akik 2007–2014 között legalább egyszer megjelentek a Szabolcs-Szatmár-Bereg Vármegyei Oktatókórházban

Ezen előfordulási gyakoriságok ugyanakkor az eddig ismert irodalmi adatoknál egy nagyságrenddel nagyobb értékeket mutattak, azaz minden millió emberre tízes nagyságrendű betegszám jutott. Ezért úgy döntöttünk, hogy klinikai és informatikai erőforrásokat felhasználva további pontosítást végzünk (5. ábra, jobb oldali panel). Ennek megfelelően, a valószínűtlen és nagyon valószínű FCS besorolást kapó, a Debreceni Egyetem Klinikai Központjában megjelent 275 beteg adatát külön ellenőrzésnek vetettük alá. Így minden elérhető részletes adatot összegyűjtve lipidológusokkal ellenőriztük a kórelőzményüket, figyelembe véve a családi halmozódásra, életkorra és a pankreatitiszre vonatkozó adatokat. Mindezek, valamint a klinikai tünetek és a TG-szintek alapján összesen 21 (7 klinikailag biztos és 14 nagyon valószínű) FCS beteget találtunk, mely az FCS prevalenciáját a Klinikai Központban megjelenő betegek körében 11,8 - 35,6 ppm-re becsüli.

Javaslatok az FCS hatékonyabb diagnosztikájához

A tanított orvosi adatokon, a gépi tanulás segítségével teljesen új összefüggések tárhatóak fel, melyek egyrészt arra használhatóak, hogy új, eddig nem diagnosztizált betegeket ismerjünk fel a sokaságban (diagnosztikai irány). A feltárt összefüggések másrészt a betegség mélyebb orvosi összefüggéseinek feltárásában is jelentős segítséget nyújthatnak (klinikai ismeretek növelése). A kutatásaink során mindkét irányban végeztünk elemzéseket, így az FH betegek esetén a gépi tanulást arra használtuk, hogy minél homogénebb betegcsoportokat képezve pontosan azonosíthassuk DEKK-nél megjelent FH betegeket, azaz a tényleges betegek tulajdonságain keresztül minél jobban megértsük az FH-ban szenvedő egyének egyéb tulajdonságait.

Az FCS tanulmányozása során, a debreceni betegek adatain, mint vektortereken tanítva, többféle gépi tanulási módszerrel matematikai modelleket építve további lehetséges FCS betegeket kerestünk mindkét egészségügyi intézményben. Tréninghalmazként a fenti 21 beteg adatát használtuk, míg teszhalmazként a fenti 275 beteg és véletlenszerűen kiválasztott egyének adatai szolgáltak. A kernel metódusok és a boosting algoritmusok itt is rendkívül jól szerepeltek. A gépi tanulás során megpróbáltuk megismerni, hogy a modellek a betegek milyen jellemző tulajdonságait veszik döntően figyelembe és hogy ezek miként alkalmazhatóak arra, hogy az FCS klinikai diagnózisát még pontosabbá tegyük.

Az elemzésekhez szükséges változókat a páciensek kórelőzményéből generáltuk és nem meglepő módon arra jutottunk, hogy FCS esetén a laboratóriumi paraméterek a

legfontosabbak. A paramétereket egy, 0-tól 100-ig terjedő skálán fontossági sorrendbe rendezték a modellek, ahol a 100-as érték jelentette a legfontosabb jellemzőt a döntéshozásban.

Megfigyeléseink megerősítették, hogy az egyes laboratóriumi paraméterek relatív fontosságának vizsgálatakor a TG szintek a legfontosabbak diagnosztikai kritériumok, hiszen az FCS is leginkább ezzel karakterizálható és a Moulin-féle pontrendszer alapján is a triglicerid szintekkel jellemezhető (13. táblázat). Kiemelt fontosságúnak találtuk az anamnézisben szereplő maximális és átlagos TG szinteket, emellett fontos volt még a TG szintek közötti fluktuáció (a legmagasabb és legalacsonyabb értékek közötti megoszlás és váltakozás) is a modellek döntésében: azaz nagyobb valószínűséggel kapott olyan egyén FCS diagnózist, ahol a kórelőzmény során a minimális és a maximális TG szintek között nem volt nagy eltérés. Fontos újdonságként megemlítendő az is, hogy a a kórelőzményben szereplő átlagos és legmagasabb koleszterin szintek szintén alapvetőnek bizonyultak abból a szempontból, hogy az FCS betegek a nem FCS páciensektől megkülönböztethetők legyenek. Ezen megfigyelésünk különösen érdekes annak tükrében, hogy az FCS diagnózisának felállítása során a koleszterin értékek vizsgálata jelenleg nem része a klinikai gyakorlatnak. A lenti táblázatban szereplő tulajdonságok így abban segítenek, hogy a valódi FCS betegeket a magas FCS pontszámmal rendelkező, de valójában a betegségben nem szenvedő egyéb hipertrigliceridemiás betegektől elkülönítsük.

13. táblázat: A kórelőzmény egyes tulajdonságainak jelentősége az FCS diagnózisában, valamennyi tanított modell eredményeinek a felhasználásával (relatív fontossági pontszámokban kifejezve, a legfontosabb jellemzők arányában, ahol a legmagasabb 100)

Igazolt és gyanított FCS betegek vs. 7+ FCS ponttal rendelkező egyének		Igazolt és gyanított FCS betegek vs. véletlenszerűen választott egyének	
Tulajdonság	Fontosság	Tulajdonság	Fontosság
Legmagasabb TG	100	Átlagos TG	100
Átlagos TG	50	Legmagasabb TG	70
Átlagos chol	25	Legalacsonyabb TG	40
TG változékonyság	20	TG változékonyság	35
Legalacsonyabb TG	17	Átlagos chol	30
Legalacsonyabb urea	16	Legmagasabb chol	25
Legmagasabb chol	15	Legalacsonyabb chol	15
Átlagos hemoglobin	14	chol változékonyság	15
Legalacsonyabb szérum glükóz	12	Átlagos hemoglobin	10
Átlagos alkalikus foszfatáz	10	Szérum glükóz változékonyság	10

További érdekes összefüggéseket találtunk, amikor megvizsgáltuk, hogy mik azok a határértékek, ahol a matematikai modellek döntésében változás következik be, azaz az FCS-ből nem FCS, illetve a nem FCS-ből az FCS diagnózisa felé mozdulnak (14. táblázat). Ezen határértékek populációs szűrésre használhatóak, de nem jelentenek abszolút fontosságot, hanem azt mutatják, hogy az adott érték alatt vagy felett az FCS diagnózisának valószínűsége nő, vagy éppen csökken. A többféle triglicerid és koleszterin “vágási” érték mellett a hemoglobin, MCHC, bazofil granulocita, limfocita vagy amiláz eredmények bemutatott értékei feletti eredmények közelebb visznek az FCS diagnózishoz, míg az emelkedett fehérvérsejt, neutrofil granulocita, nátrium, glükóz, CRP, ill. máj- és vesefunkciós értékek pedig csökkentik az FCS diagnózis valószínűségét. Érdekes, hogy emelkedett gyulladásos paraméterek, így magasabb fehérvérsejtszám és CRP jelenléte szintén csökkenti az FCS diagnózis lehetőségét – legalábbis jelen tanulmányunkban.

14. táblázat: A leginkább meghatározó laboratóriumi paraméterek határértékei az FCS diagnózisában. (+): FCS valószínűségét növeli, (-): FCS valószínűségét csökkenti

Vizsgálat	Határérték (>)	Hatásirány
Triglicerid	30 mmol/l	+
Triglicerid	18 mmol/l	+
Triglicerid	6.5 mmol/l	+
Koleszterin	11 mmol/l	-
Koleszterin	6.5 mmol/l	+
Koleszterin	4.0 mmol/l	+
Hemoglobin	95 g/l	+
MCHC	330 g/l	+
Bazofil granulocita arány	0.6%	+
Limfocita arány	20%	+
Amiláz	20 U/l	+
Fehérvérsejtszám	6.5 G/l	-
Neutrofil granulocita arány	65%	-
Nátrium	145 mmol/l	-
Glükóz (éhom)	6,0 mmol/l	-
CRP	5,0 mg/l	-
GPT	15 U/l	-
GPT	200 U/l	-
GGT	35 U/l	-
GGT	350 U/l	-
Kreatinin	68 µmol/l	-

FH betegek komplex vizsgálata Lp(a) kategóriák szerint

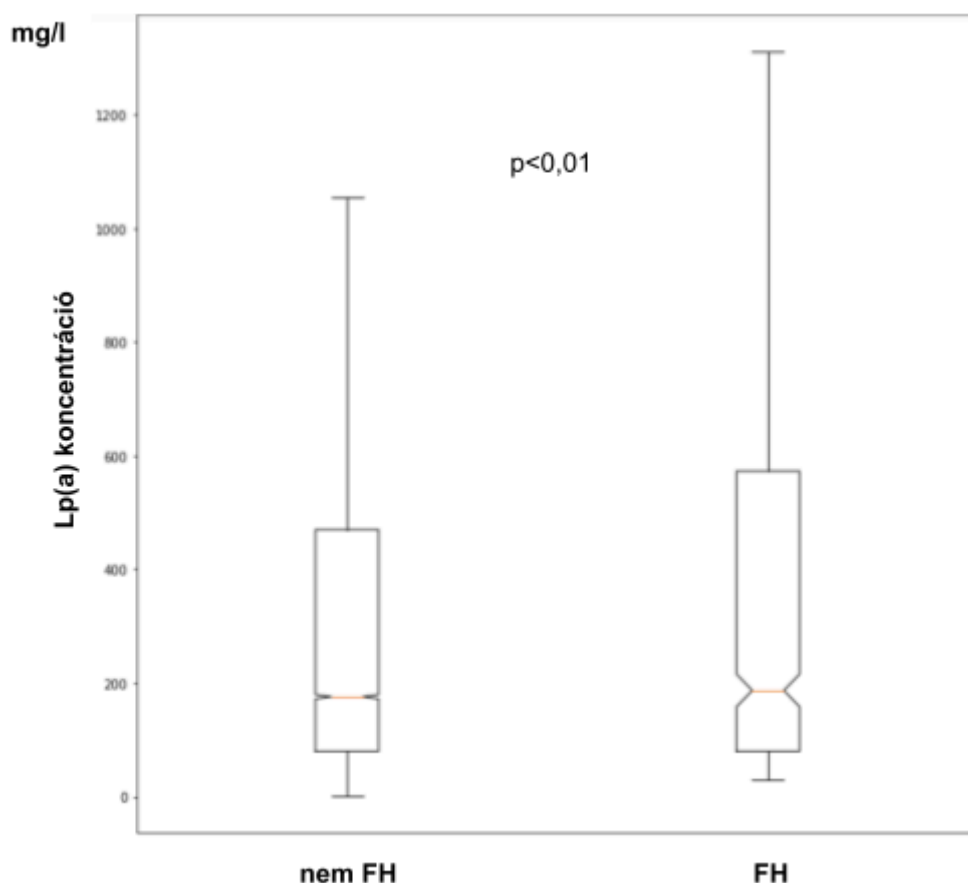
A ritka kórképek kiemelt tulajdonságainak feltárásához, valamint a más klinikai állapotokkal esetlegesen létező összefüggések feltárásához fontos a meglévő adatokhoz való gyors kapcsolódás, valamint az egészségügyben történő megjelenések során keletkező klinikai rekordok tulajdonságainak kiemelése (pl. a szöveges adatok kategorizálása), melyeket az “Anyag és Módszer” fejezetben részleteztünk. Ezek során az FH betegeknél részletes eredményeket kaptunk a kardiovaszkuláris rizikó, életmód és laborértékek vonatkozásában. Utóbbiakból is kiemelendő az Lp(a) szint, mely fontos független kardiovaszkuláris rizikófaktor, ugyanakkor az FH betegek nagy kardiovaszkuláris morbiditásához és mortalitásához való hozzájárulása mind a mai napig intenzíven kutatott terület. Az Lp(a)-val foglalkozó kutatásainkat kizárólag a Debreceni Egyetem Klinikai Központjában megjelent betegek (n=590.500) körében végeztük. Az adatbányászat és gépi tanulási módszerek alkalmazását követően a legjobb tréning eredményeket a DLCN pontszám 7 és 8 feletti értékeinél kaptuk, mely alapján – a DLCNS koncepciójával összhangban – a 7, vagy afeletti pontszámmal rendelkező egyéneket FH betegnek tekintettünk. Mindezek alapján 459 beteg volt azonosítható, a vizsgált populációt (FH és nem FH betegek) a 15. táblázatban mutatom be.

15. táblázat: A vizsgált populáció klinikai és laboratóriumi adatai. Az adatok mediánként (alsó-felső kvartilis) vannak megadva, n.s.: nem szignifikáns eltérés az FH és nem FH csoport között

	FH	nem FH	p
Összes beteg	459	590.500	
Életkor (év)	53,2 (39,3-59,7)	40,0 (22,2-59,0)	<0,05
Férfi/nő (%)	39,2/60,8	42,8/57,2	n.s.
CHOL (mmol/l)	7,03 (6,10-8,40)	4,94 (4,21-5,73)	<0,001
TG (mmol/l)	1,96 (1,41-3,03)	1,35 (0,96-1,96)	<0,01
LDL-C (mmol/l)	4,80 (3,91-5,90)	2,95 (2,34-3,62)	<0,001
HDL-C (mmol/l)	1,41 (1,16-1,72)	1,32 (1,09-1,62)	<0,05
Urea (mmol/l)	5.43 (4.40-6,63)	5.10 (4,02-6,50)	n.s.
Kreatinin (μmol/l)	69,1 (58,5-82,2)	68,3 (57,1-83,2)	n.s.
eGFR (ml/min/1,73m2)	88,8 (80,4-90,0)	90,0 (78,1-90,0)	n.s.

GOT (U/l)	21,8 (18,3-27,5)	21,0 (17,5-28.1)	n.s.
GPT (U/l)	24,0 (16,5-33,0)	19,6 (14,0-29.2)	n.s.
GGT (U/l)	32,7 (20,2-58,9)	26,1 (16,1-48.2)	<0,05
Húgysav (μmol/l)	324 (266-384)	292 (235-357)	<0,05
Lp(a) (mg/l)	236 (92,5-698,5)	167 (80,2-431,5)	<0,01

Az azonosított 459 FH beteg közül 221 esetén (48%) volt elérhető Lp(a) eredmény. Az FH betegek szignifikánsan magasabb Lp(a) értékekkel rendelkeztek, mint a nem FH-s betegpopuláció [236 (92,5; 698.5) vs. 167 (80,2; 431.5) mg/l, $p < 0,01$] (6. ábra). Mindkét csoportban azt találtuk, hogy a nőknek szignifikánsan magasabb Lp(a) értékeik voltak, mint a férfiaknak [266 (108–875) vs. 182 (73,1–648) mg/l, $p < 0,05$] az FH betegek között és [179 (81,5–458) vs. 152 (80,7–386) mg/l, $p < 0,01$] a nem FH egyének körében. Bár az ateroszklerotikus szövödmények lényegesen gyakoribbaknak bizonyultak az FH betegekben, mint a nem FH csoportban (46,6% vs. 13,9%), ugyanakkor – a korábbi irodalmi adatokkal ellentétben – az Lp(a) szintek és az érlemezsedéses szövödmények gyakorisága között nem találtunk szignifikáns összefüggést az FH csoporton belül.



6. ábra. A szérumban Lp(a) szintek a két csoportban (FH betegek és nem FH-s egyének). A box hossza az interkvartilis tartományt (IQR), a vízszintes vonal a doboz belsejében a mediánt, a whiskerek a 25. kvartilis 1,5 IQR-jét vagy a 75. kvartilis 1,5 IQR-jét jelzik.

Utóbbi megfigyelésünkre való tekintettel, ezt követően megvizsgáltuk az egyéb kardiovaszkuláris kockázati tényezők előfordulási gyakoriságát mindkét csoporton belül (16. táblázat).

16. táblázat: A kardiovaszkuláris kockázati tényezők prevalenciája FH és nem FH betegekben, valamint különböző ateroszklerotikus eredetű megbetegedésekben és mélyvénás trombózisban

	FH	nem FH	CVD	CeVD	PAD	AoS	DVT
magas Lp(a) (%)	35,3	23,0	25,7	25,8	23,8	33,9	26,2
hipertónia (%)	86,3	24,5	77,8	82,6	61,1	85,0	64,2
diabétesz (%)	17,4	5,16	19,3	24,4	14,7	23,0	15,7
dohányzás (%)	66,4	9,25	28,0	32,2	30,6	28,7	25,0
elhízás (%)	42,0	10,2	31,4	36,1	26,3	33,5	38,2
hiperurikémia (%)	41,2	5,78	19,2	24,4	21,0	22,1	24,7
alacsony HDL-C (%)	11,8	2,12	8,59	9,23	6,17	9,79	7,96

férfiak (%)	39,2	42,7	48,1	46,5	26,6	57,8	42,4
60 év feletti életkor (%)	20,0	23,0	54,8	64,7	33,7	77,8	50,5
nincs statin kezelés (%)	12,9	88,3	46,6	45,3	56,2	35,6	68,8

Rövidítések jelentései: CVD (cardiovascular disease, szív és érrendszeri betegségek), CeVD (cerebrovascular disease, agyi érrendszeri megbetegedés), PAD (peripheral arterial disease, perifériás artériás betegség), AoS (aortic valve stenosis, aortabillentyű-szűkület), DVT (deep vein thrombosis, mélyvénás trombózis).

Eredményeink alapján különösen a magasvérnyomás-betegség és a dohányzás, de az elhízás, emelkedett húgysavszint, a cukorbetegség és az alacsony HDL koleszterinszint is sokkal gyakoribb az FH betegek körében. Ahhoz, hogy tisztázzuk az Lp(a) szerepét a betegséggel érintett és az attól mentes populációban, mélyebben megvizsgáltuk az ateroszklerotikus eredetű betegségek és a kardiovaszkuláris kockázati tényezők kapcsolatát. Az eredmények alapján a legfontosabb kockázat egyértelműen a magas vérnyomás, a 60 év feletti életkor és a statin kezelés hiánya, valamint a dohányzás és az elhízás volt. Szintén jelentős kockázatnak számított a férfi nem. Ez csaknem minden vaszkuláris szövődmény esetén hasonlóan bizonyult (CVD, CeVD, PAD, DVT, AoS). A legmagasabb arányú emelkedett Lp(a) értéket az AoS betegeknel találtunk.

A kockázati tényezőket magas és alacsony Lp(a) értékkel rendelkező FH betegeknel is megvizsgálva megállapítható, hogy lényeges eltérés nincs a két csoport között (17. táblázat). A nemek közötti összehasonlításban ugyanakkor megfigyelhető, hogy magas Lp(a) értékek minden csoportban gyakrabban fordulnak elő nők esetében, bár az FH betegek esetén az eltérés még így sem szignifikáns. Emellett a nem FH-s férfiak körében a dohányzás és hipertónia gyakorisága alacsonyabb volt.

17. táblázat. A kardiovaszkuláris kockázati tényezők prevalenciája a teljes vizsgálati populációban, FH betegekben, illetve alacsony és magas (>500 mg/dl) Lp(a) szinttel rendelkező FH betegekben.

A teljes vizsgálati populáció				
	nem-FH	FH	FH + alacsony Lp(a)	FH + magas Lp(a)
magas Lp(a) (%)	23,0	35,3	-	-
magas vérnyomás (%)	24,5	86,3	86,8	87,2
diabétesz (%)	5,16	17,4	16,4	12,8
dohányzás (%)	9,25	66,4	69,7	76,9
obezitás (%)	10,2	42,0	42,8	37,2
hyperuricemia (%)	5,78	41,2	40,8	46,2
alacsony HDL-C (%)	4,63	31,4	31,6	28,3
férfiak (%)	42,7	39,2	32,8	27,3

60 év feletti kor (%)	23,0	20,0	24,7	27,1
nincs statin kezelés (%)	88,3	12,9	16,4	9,0
Férfiak				
	nem-FH	FH	FH + alacsony Lp(a)	FH + magas Lp(a)
magas Lp(a) (%)	21,4	29,4	-	-
magas vérnyomás (%)	25,8	86,7	87,8	87,8
diabétesz (%)	5,94	20,1	20,4	14,7
dohányzás (%)	11,9	64,5	69,4	78,7
obezitás (%)	9,71	46,1	51,0	39,0
hyperuricemia (%)	5,64	43,9	46,9	49,1
alacsony HDL-C (%)	4,58	28,8	30,6	25,4
férfiak (%)	-	-	-	-
60 év feletti kor (%)	23,2	17,2	25,1	27,3
nincs statin kezelés (%)	86,6	11,7	18,4	9,20
Nők				
	nem-FH	FH	FH + alacsony Lp(a)	FH + magas Lp(a)
magas Lp(a) (%)	24,1	37,9	-	-
magas vérnyomás (%)	23,6	86,0	86,4	86,2
diabétesz (%)	4,59	15,8	14,6	12,1
dohányzás (%)	7,27	67,7	69,9	75,9
obezitás (%)	10,6	39,4	38,8	36,2
hyperuricemia (%)	5,88	39,4	37,9	44,8
alacsony HDL-C (%)	4,67	33,0	32,0	29,3
férfiak (%)	-	-	-	-
60 év feletti kor (%)	22,9	20,9	24,1	26,9
nincs statin kezelés (%)	89,3	13,6	15,5	8,66

Hogy átfogóbb képet kapjunk az egyes kockázati tényezők önmagukban értelmezhető kockázatáról, a nyers adattömegben meghatároztuk a kockázati tényezőkkel összefüggésbe hozható kardiovaszkuláris megbetegedések relatív házárd értékeit is (HR, hazard ratio). Ehhez első lépésben definiálni kellett egy “egészséges” populációt, melyhez képest az egyes kockázati tényezők jelentőségét vizsgálni tudtuk. Nyilvánvaló, hogy “egészségesnek” itt a következőkben részletezett kondícióktól mentes egyéneket tekintettük, de más betegségeik lehettek. Az “egészséges” populációba olyan egyének tartoztak, akiknél túlsúly, dohányzás,

sztatinkezelés, hipertónia, diabétesz, magas Lp(a) nem volt megállapítható, emellett a normál tartományba eső LDL-C és HDL-C értékekkel rendelkeztek (18.táblázat).

18. táblázat: Az egyes kockázati tényezők hatása a kardiovaszkuláris rizikókockázatra (relatív házárdok)

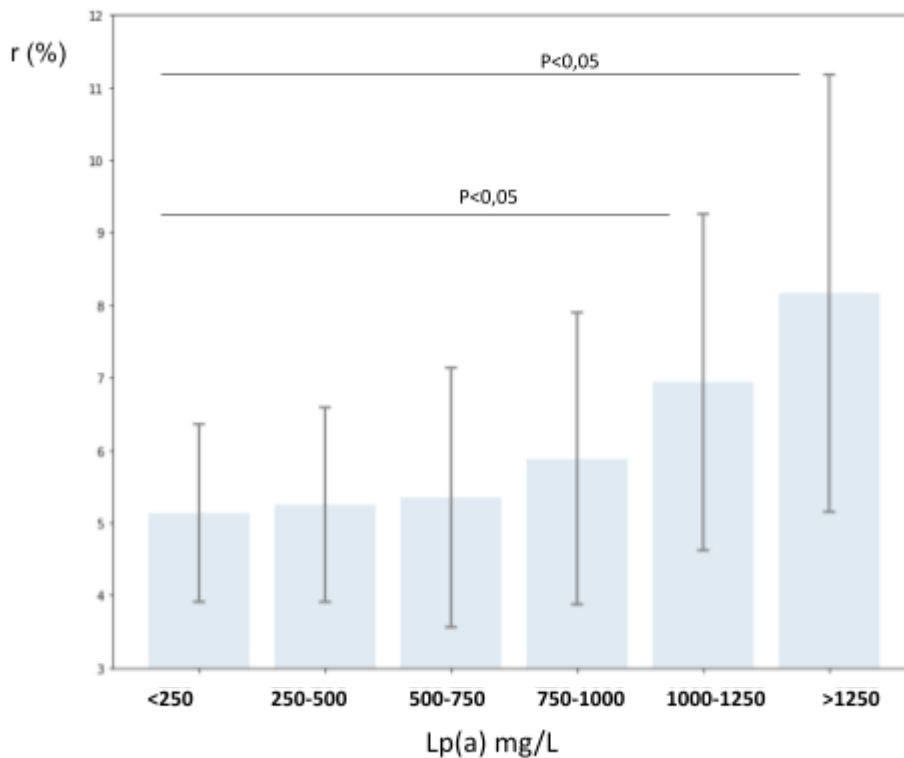
	HR (egészséges*)	HR (egészséges FH**)
dohányzás	4,9 – 5,2	1,3 – 10,6
elhízás	1,9 – 2,2	-
hipertónia	10,3 – 10,4	4,3 – 9,1
diabétesz	2,3 – 2,7	-
magas LDL-C	11,3 – 11,9	-
extrém magas Lp(a) (>1000 mg/l)	1,6 – 6,4	-

*Az alábbi szempontokból teljesen egészségesnek tekinthető, azaz nem túlsúlyos, nem dohányzó, sztatinkezelés nélküli, normális LDL-C és HDL-C szinttel rendelkező, normotoniás, cukorbetegségtől mentes és normális Lp(a) értékekkel rendelkező betegek relatív házárd értéke (392.600 beteg)

** FH-betegek, akik nem túlsúlyosak, nem dohányoznak, nem szenvednek magasvérnyomás-betegségben, vagy cukorbetegségben és nem rendelkeznek magas Lp(a) értékkel (39 beteg)

Azt találtuk, hogy a legmagasabb relatív házárd értékkel a magasvérnyomás-betegség és magas LDL-C értékek rendelkeznek, melyeket szorosan követett a dohányzás. FH betegekben mind a hipertónia, mind pedig a dohányzás jelentősen növeli a már amúgy is magas kardiovaszkuláris kockázatot. A betegek jelentős része ugyanakkor egynél több kockázati tényezővel is rendelkezett. Azoknál a betegeknél, akik az összes kockázati tényezővel halmozottan rendelkeztek (leszámítva a magas Lp(a) kockázatot, melynek ilyen esetben a relatív házárdja már nem volt értelmezhető) a relatív házárd 27,42-nek adódott: azaz egy egészséges emberhez képest a kockázati tényezőket többszörösen hordozó beteg kardiovaszkuláris kockázata közel 30-szoros a rizikófaktor nélküliekhez képest.

Végül az Lp(a) szinteket az egész populációban részletesen, a többi kockázati tényezőtől függetlenül elemezve arra jutottunk, hogy az 1000 mg/l feletti szérumszintek esetén a kardiovaszkuláris rizikó szignifikánsan nő. Ezen határérték alatt – bár fokozatos növekedés észlelhető – a rizikó fokozódás, így az egyén élete során bekövetkező kardiovaszkuláris események gyakoriságának emelkedése nem szignifikáns (7. ábra).



7. ábra: Az élet során bekövetkezett kardiovaszkuláris események aránya (%) a különböző szérumban Lp(a) szinttel rendelkező csoportok szerint kategorizálva, ismert Lp(a) értékkel rendelkező FH és nem FH betegeknél

Saját munkám

Az alábbi folyamatokban járultam hozzá a kutatócsoport munkájához, a fenti eredmények eléréséhez:

- Adatbányászati és gépi tanulási módszerek megtervezése;
- Az egészségügyi adatok tisztítása, előfeldolgozása, illetve a lekérdezésekhez és a gépi tanuláshoz alkalmas formátumra alakítása, beleértve a szkriptek megírását, az adatbázisok beállítását és paraméterezését;
- Adatbányászati és gépi tanulási módszerek kidolgozása, maguknak a kódoknak a megírása és tesztelése;
- Adattáblák közötti lekérdezések, statisztikák, másodlagos adatok (például kockázati tényezők) és diagramok készítése, valamint az ezekhez szükséges irodalmi kontextus megértése, irodalmazás elvégzése;
- Az előszűrt kórtörténeti adatok előzetes orvosi szempontok szerinti átnézése, osztályozása és kigyűjtése;
- A fenti munkafázisokból származó adatok értékelése.

Új eredmények összefoglalása

1. A DLCNS pontrendszer adatbányászati eszközökkel történő alkalmazása során a hazai regionális FH prevalenciát 1:340-nek határoztuk meg. Gépi tanulási módszerek alkalmazásával a biztosan FH-s betegek aránya 1:1300.
2. A korábban javasolt pontrendszer adatbányászati módszerekkel történő alkalmazásával az FCS regionális gyakorisága 19,4:1.000.000-ra; mely a két egészségügyi szolgáltatót külön vizsgálva 13,3 - 27,11 ppm közötti értékre becsülhető. Gépi tanulási módszerekkel az FCS prevalenciáját 11,8 - 35,6 ppm-re becsültük, mely prevalencia az eddig ismert irodalmi adatok körülbelül 10-szerese.
3. Ensemble módszerekkel és sekély neurális hálózatok segítségével hatékony matematikai modelleket fejlesztettünk, melyek alkalmasak rutin vérvételi eredmények segítségével elkülöníteni az FCS betegeket az egyéb klinikai betegektől (átlagos AUC: 0,98); ill. elkülöníteni az FCS betegeket az egyéb, hipertrigliceridémiában szenvedő nem FCS betegektől (átlagos AUC: 0,87).
4. Az általunk fejlesztett gépi tanulás modellek részletes vizsgálata alapján javasolhatóak olyan laborparaméterek, melyek alkalmasak lehetnek az FCS pontrendszer további finomítására. Ezek közül a szérum össz-koleszterin szintje kiemelten jó jelölt, hiszen a magas értékek javítják az FCS diagnózisának esélyét, míg a nagyon magas értékek csökkentik azt. Így meghatározható a jövőben egy sáv, ahol az FCS diagnózis valószínűsége maximális (jelen vizsgálataink alapján ez a sáv 6,5 - 11,0 mmol/l között van).
5. Az FH betegek részletes vizsgálata során az ateroszklerotikus kardiovaszkuláris kockázat lényegesen nagyobb voltát igazoltuk FH betegekben (46,6% vs. 13,9%), ahogy az Lp(a) szérumértékek is lényegesen magasabbak voltak az FH-tól mentes populációhoz képest [236 (92,5; 698,5) mg/l vs. 167 (80,2; 431,5) mg/l]. A magasabb Lp(a) értékek és a kardiovaszkuláris rizikó emelkedése között közvetlen összefüggést nem tudtunk igazolni FH betegek esetén. Ugyanakkor, az összes kardiovaszkuláris kockázati tényező előfordulási aránya lényegesen magasabb volt az FH betegek esetén, míg az egyéni relatív házárd értékek azonosak.
6. A nemi különbségeket tekintve, a szérum Lp(a) szintek a nők között FH-ban és FH-val nem érintett populációban is szignifikánsan magasabbnak bizonyultak.
7. Kutatásaink alapján az 1000 mg/l feletti Lp(a) értéknél a kardiovaszkuláris rizikó szignifikánsan megnő. Az 1000 mg/l feletti Lp(a) érték relatív házárdja 1,6-6,4

közötti, mely bár jelentős, de mégis alacsonyabb a gyakori másodlagos tényezők (így a dohányzás, vagy a hipertónia) egyéni kockázatához képest.

8. Eredményeink igazolják, hogy a nagy mennyiségű klinikai adaton alkalmazott adatbányászati és gépi tanulási módszerek hatékonyak lehetnek a mindennapos klinikai munka, diagnosztika és tudomány támogatásában.

Megbeszélés

Ritka betegségek prevalenciája regionális és globális összehasonlításban

Bár a ritka kórképek kevesebb egyént érintenek a populációban, ugyanakkor ezen személyek potenciális egészségkárosodása, illetve az ebből eredő orvosi ellátás költsége mind a betegre, mint a társadalomra jelentős egészségügyi terhet képes róni. Jellegükből adódóan, ezen kórképek pontos gyakoriságának meghatározása is igen nagykihívást jelent. A genetikai prediszpozíciótól függően, a familiáris hiperkoleszterinémia 1:250 - 1:1.000.000 gyakorisággal fordul elő a populációban [5], míg a familiáris kilomikronémia szindróma prevalenciája szintén 1:1.000.000-ra tehető [54]. Az FH betegek esetén az ateroszklerotikus eredetű kardiovaszkuláris megbetegedések jelentik a fő halálokat; az FCS esetén pedig – az ASCVD mellett – a pankreatitisz és annak szövődményei befolyásolják jelentősen az életminőséget és a várható élettartamot [6]. Bár gyakoriságukból és súlyosságukból adódóan ezen betegségek Magyarországon és régiókban is jelentős kihívást jelentenek a szakemberek számára, ugyanakkor az FH és FCS pontos prevalenciája nem ismert.

Tudomásunk szerint hazánkban elsőként alkalmaztuk az adatbányászat és gépi tanulás módszereit nagy mennyiségű egészségügyi adaton ritka betegségek prevalenciájának becsléséhez. A DLCNS pontrendszer alkalmazásával végzett kalkuláció alapján az észak-alföldi régióban minden 1.000.000 lakosra 340 FH beteg juthat, ami több európai ország adataival is összhangban van [55]. Az FH szűrésének kiemelt szerepe van a kardiovaszkuláris megbetegedések megelőzésében, emellett az újonnan fejlesztett és hatékony lipidcsökkentő terápiák (pl. a PCSK9-gátlók) szintén hozzájárulhatnak ahhoz, hogy az FH szűrések felgyorsuljanak, mivel egyre több beteg helyzete javítható ezekkel az új terápiás lehetőségekkel [56]. Magyarország szintén csatlakozott a legtöbb, FH szűrésére irányuló nemzetközi programhoz és a Debreceni Egyetem Belgyógyászati Intézet Anyagszere Betegségek Tanszékének vezetésével Debrecenben is FH regiszter indult a régióban. Az erőfeszítéseink ellenére a betegek regisztrálása lassúnak bizonyult, ezért is fordultunk egyéb módszerek felé a betegek lehetséges azonosításában, a régió két legnagyobb forgalmú egészségügyi intézményének nyolc teljes évet felölő adatát áttekintve. Bár a mintavétel nem tekinthető minden szempontból reprezentatívnak, mégis a régió lakosságának kb. $\frac{3}{4}$ -ét

érintette, a lakosság kórházba járási szokásait figyelembe véve. A debreceni betegek adatain később elvégzett gépi tanulási módszerekkel tovább finomítottuk az FH betegek azonosítását és 459 valószínűsíthetően FH beteget azonosítottunk, mely némileg csökkentette az FH prevalenciáját 1:1300-ra. Fontos megemlíteni, hogy a cél utóbbi esetben az volt, hogy az azonosított egyén biztosan FH beteg legyen – így némileg alulbecsülhettük a tényleges prevalenciát (bővebben a kutatás korlátairól a fejezet végén).

Az FCS-t hasonló módszerekkel vizsgálva a betegség gyakoriságát 19,4 ppm-nek találtuk, ami nagyobb, mint a becsült 1 ppm globális prevalencia [54]. Ugyanakkor az FCS a mi eredményeink alapján is ritka betegségnek számít, mivel definíció szerint minden kórkép, amely előfordulása az 1:2000-et nem haladja meg, ritka betegségnek tekintendő. Megemlítendő, hogy a frissebb tanulmányok a mi eredményeinkkel összevethető adatokkal szolgáltak. Így pl. mintegy másfél millió beteg adatát áttekintve, az FCS prevalenciája 13 ppm-nek adódott egy, a mi vizsgálatainkhoz hasonló, egészségügyi szolgáltatók által nyújtott ellátást igénybevevő populáción [57]. Gyermekekben, kisebb betegtömeget nézve ez a szám 1:300.000-nek adódott [58], ami szintén magasabb prevalenciát sejtet, mint a jelenleg általánosnak elfogadott becslések.

Itt fontos azt is megemlíteni, hogy a prevalenciában tapasztalható eltérések részben a mintavétellel is magyarázhatóak attól függően, hogy egészséges, vagy beteg (pl. klinikai, vagy kórházi) populációból vesszük a mintát. Befolyásoló tényező lehet még a vizsgált intézmény ellátási szintje is, mivel a kiemelt szintű betegellátást végző centrumokban a becsült prevalencia is nagyobb lehet, hiszen az alaposabb kivizsgálást, szélesebb körű laboratóriumi vizsgálatokat igénylő betegek nagyobb arányban jelennek meg leginkább specializált ellátási intézményekben. Hasonló eredményre jutottunk saját elemzéseinkkel is, akár a kulcsjellemzők alapján végzett gyors becslést, akár pedig az egyes betegeket leíró, egyedi pontszámokat tekintve.

Vizsgálataink alapján mindenesetre feltehető, hogy az FCS becsült prevalenciája a jelenleg elfogadottnál egy nagyságrenddel nagyobb lehet, melyet az általunk, nagy betegtömegeken alkalmazott módszertan (adatbányászat és gépi tanulás) is alátámaszt. Utóbbi azért is fontos, mivel az FCS diagnosztikája jelenleg nagyban támaszkodik a csak specializált központokban elérhető a genetikai analízisre, valamint az LPL enzimaktivitás mérésére [15]. A közelmúltban ugyanakkor fontos változást hozott a lipidológus szakértők által alkotott FCS score, mely az FCS gyors és hatékony felismerését támogatja [14]. Az FCS pontrendszer alapvetése a szérum triglicerid szint ismerete, mely TG-szint minden ellátási szinten könnyen mérhető. Fontos azt is megemlíteni – ahogy mi is igazoltuk kutatásaink során –, hogy csupán

a szérumban TG szintjével, illetve annak meghatározott értékeinek vizsgálatával (pl. előfordult legalább háromszor 10 mmol/l feletti TG-szint és sosem volt 2 mmol/l alatt) a tovább vizsgálható populáció 1-2%-ra szűkíthető. Amennyiben egy újabb erős és könnyen mérhető, szubjektivitástól mentes paraméterrel bővítjük az analízist (így pl. a diagnosztikailag elvárt TG értéket a beteg életében legalább egy alkalommal 20 mmol/l-re emeljük), akkor a vizsgálható betegek száma az eredeti populáció tízezrelékére esik úgy, hogy közben elhanyagolható eséllyel vétünk fals negativitásból eredő hibát.

Ez a szoros és nyilvánvaló kapcsolat a szérumban TG szintek és a betegség megléte között lehetővé teszi egyszerű "score" rendszerek sikeres alkalmazhatóságát a gyakorlatban, ugyanakkor azt találtuk, hogy a ténylegesen FCS-sel diagnosztizált betegek sokszor nem azok, akik a legmagasabb FCS score-t kapták. Ez magyarázható azzal is, hogy nem rendelkezünk mindazzal az anamnesztikus adattal, amelyek szükségesek a pontszám biztos kiszámításához. Ezek nem is feltétlenül érhetők el minden ellátási szinten, de felhívják a célzott és részletes kérdésekkel felvett kórtörténet jelentősége is a figyelmet: így pl. nagyon fontos az összes másodlagos, hipertrigliceridémiát provokáló tényező kizárása is az anamnézis felvétele során [59]. Fentiek figyelembe vételével úgy gondoljuk, hogy a korábban javasolt FCS pontrendszer kiemelkedő hatékonysággal működik minden esetben, ahol teljeskörűen elérhető minden adat a páciensről, viszont korlátozottabb (és tovább pontosítható) minden olyan esetben, ahol ez a feltétel nem áll fenn. Ebben viszont a gépi tanulás és adatbányászat hatékony kiegészítésül tud szolgálni, különösen ha nagy adathalmazokban, automatizált módszerekkel próbáljuk a potenciális FCS betegeket azonosítani, mely az új gyógyszerfejlesztések és kezelési lehetőségek tükrében egyre fontosabb és jelentősebb feladat [43,60].

Klinikai kritériumok és protokollok finomítása

A klinikum támogatásához, így a pontos, jól használható protokollok kialakításához és az FCS patomechanizmusának jobb megértéséhez is hasznos eszköznek tűnik a gépi tanulás. Intézetünk lipidológusai segítségével, a diagnosztizált FCS betegek adatain tanítva azt találtuk, hogy a bináris klasszifikáció során a kernel és boosting alapú modellek nagyon hatékonyan, magas ROC/AUC értékek mellett voltak képesek az FCS és nem FCS betegeket szétválasztására, ugyanakkor a neurális hálók lényegesen rosszabbul teljesítettek. A

laboratóriumi paraméterek részletesebb vizsgálata során egyrészt az egyes paraméterek fontosságát próbáltuk becsülni a modellek predikciója során. Így bizonyos paraméterek elvételével és hozzáadásával újabb tanításokat végeztünk, másrészt a már sikeresnek bizonyult modellekbe változó laborértékeket táplálva próbáltuk megtalálni a legfontosabb értékeket (ún. “vágásokat”), ahol a “beteg - nem beteg” bináris szeparáció a legjobb hatékonysággal végbemegy. Azt találtuk, hogy a szérumban TG értékek vizsgálata nyilvánvalóan segíti – és az összes laboratóriumi érték közül a legjobban segíti – az FCS pontos diagnózisát. Ennek megfelelően, az extrém magas (így a 20 mmol/l-t meghaladó) értékek, valamint a közepesen magas, de ritkán változó TG szintek tovább pontosítják a diagnózist.

Véleményünk szerint érdekes és fontos újdonság, hogy a szérumban összkoleszterin értékeinek vizsgálata további segítséget tud nyújtani a lehetséges FCS betegek megtalálásában, mely koleszterin szintek értékelése az eredeti FCS score rendszerben jelenleg szerepel. Vizsgálataink alapján a 4 mmol/l alatti és a 11 mmol/l feletti koleszterin értékkel rendelkező páciensek esetén az FCS diagnózisának valószínűsége kicsi, mely a triglicerid és koleszterin metabolizmusának, illetve a hordozó lipoproteinek szoros kapcsolatára is felhívja a figyelmet [61].

A lipidparaméterek mellett számos egyéb metabolikus komponenst is találtunk (pl. a máj transzamináz aktivitások, vagy a szérumban glükóz-koncentráció), melyek emelkedett volta negatívan befolyásolta az FCS valószínűségét. Ezen megfigyeléseink valószínűleg az egészségügyi szolgáltatóknál megjelenő betegek körében gyakori inzulinrezisztencia, ill. az ehhez kapcsolódó kórállapotok, így az elhízás, 2-es típusú diabétesz és a nem-alkoholos zsírmáj betegség (non-alcoholic fatty liver disease, NAFLD) rendszeres előfordulásával magyarázhatóak. Ez egyébként összhangban van egy, néhány éve megjelent tanulmány megállapításaival is, miszerint szignifikánsan magasabb transzamináz aktivitások észlelhetők MFCS-ben, mint FCS-ben [15]. Megfigyeléseinkkel egybevág az is, hogy az NAFLD lényegesen gyakrabban fordul elő MFCS betegekben, mint FCS páciensek esetén [62]. Mivel a NAFLD ismert kardiovaszkuláris rizikófaktor, kutatásaink így egyrészt segíthetnek az FCS betegek kardiovaszkuláris kockázatának pontosabb feltárásában, másrészt hozzájárulhatnak az FCS lefolyásának pontosabb feltárásához. Úgy gondoljuk, hogy fenti eredményeink az FCS iránti tudatosságot, a diagnosztikai gondolatmenetbe való beépülését is elősegíthetik és mindenképpen felhívják a figyelmet olyan egyéb kórállapotokra, laboreltérésekre, melyek inkább a multifaktoriális eredetet valószínűsítik és csökkentik az FCS diagnózisának valószínűségét.

FH betegek patogenezise a metabolikus eltérések és egyéb tényezők tükrében

A nagy adathalmazokon végzett matematikai-statisztikai műveletek alkalmasak arra is, hogy segítségükkel a betegségek kórlefolyásának számos tényezőjét vizsgáljuk, majd azok között összetett függőségi és oksági viszonyokat tárjunk fel és ezeket jól definiált csoportok között összehasonlítsuk. Ilyen elemzésre, így pl. metabolikus paraméterek eltérő szintjeinek hatásvizsgálatára ad lehetőséget a szérumban Lp(a) szintje és az FH betegek kardiovaszkuláris kockázata közötti összefüggések vizsgálata. Bár az FH betegek mindegyikének magas az LDL-C szérumszintje, mégis – széles fenotípusos heterogenitást mutatva – különböző gyakorisággal és dinamikával jelentkeznek náluk a kardiovaszkuláris megbetegedések [63]. Ismert, hogy a különböző kockázati tényezők együttes előfordulása jelentősen növeli az ateroszklerózis kockázatát, hozzájárulva a fenotípusos heterogenitáshoz mind az FH, mind pedig a nem FH betegek körében. Korábbi vizsgálatok az arcus corneae, valamint a cukorbetegség és a dohányzás szignifikáns szerepét írták le az FH betegek kardiovaszkuláris kockázatának emelésében [64]. Bár már azonosították az Lp(a) szintje, a carotis szűkülete valamint a CVD kockázat közötti kapcsolatot heterozigóta FH páciensekben [65], az emelkedett Lp(a) koncentrációk rizikófaktoraként való potenciális szerepe mind a mai napig nem tisztázott ezen betegségben. Emellett megemlítené az is, hogy számos tanulmány vizsgálta az FH és a különböző CVD kockázati tényezők kapcsolatát – leginkább vegyes, sokszor ellentmondó eredményekkel.

Ezen korábbi tanulmányok tükrében, hazánkban elsőként megpróbáltuk a magyar FH betegek kardiovaszkuláris kockázati tényezőit adatbányászati módszerekkel feltérképezni. Az általunk vizsgált biztosan FH-ban szenvedő populációban a hipertóniás betegek aránya extrém magasnak bizonyult (86,3%), amely hasonló egyéb kelet-közép-európai adatokhoz [66], [62] [68]. Ugyanakkor, nyugat-európai és dél-amerikai populációkon végzett felmérések során a magasvérnyomás-betegség prevalenciája ennél lényegesen kisebbnek, 11-24% közöttinek adódott FH betegekben [69], [65], [71].

Hasonló módszerekkel megvizsgálva a dohányosok arányát, azt találtuk, hogy hazánkban meglepően sok FH beteg dohányzik (66,4%), melyhez igen hasonló adatot találunk a lengyel betegek körében [66], míg más országokban ez az arány lényegesen kisebb volt [68],[67],[65], [67], [64]. A cukorbetegség tekintetében az általunk vizsgált FH kohort nagyon hasonló volt a lengyel, román vagy a mexikói FH betegekhez [61, 63, 65], így a

diabétész gyakorisága lényegesen nagyobbak bizonyult a betegek körében, mint a non-FH populációban. Ezzel ellentétben, Vohnout és mtsai. szlovák FH betegeket vizsgálva, körükben kisebb diabétész prevalenciát tapasztaltak, mint a non-FH egyéneknél (10.5% vs. 17.4%) [73]. Ezen, változatos adatok is aláhúzzák a rizikófaktorok vizsgálatának fontosságát, valamint rávilágítanak a széleskörű, nemzetközi felmérések szükségességére FH betegeknél. Érdekes megjegyezni azt is, hogy vannak tanulmányok, amelyek sokkal kisebb diabétész prevalenciát tapasztaltak és olyan megfigyelést is tettek, mely szerint a mutációk következményének súlyossága FH betegek esetén fordítottan arányos a diabétész prevalenciával [74]. Ezeket az eredményeket nekünk nem sikerült megerősíteni. Úgy véljük, hogy az ellentmondó adatokat a páciensek változatos társadalmi és gazdasági státusza, valamint a genetikai eltérések is eredményezhetik. Emellett a páciensek és hozzátartozóik legtöbbször hiányos tudással rendelkeznek mind az FH, mind pedig az általa képviselt kardiovaszkuláris kockázat tekintetében, ami szintén hozzájárulhat a módosítható rizikófaktorok nagyon magas előfordulási gyakoriságához, beleértve a dohányzást is. Ez különösen jellemző a magyar FH betegek esetén, de a fentiek alapján látható, hogy talán ennél nagyobb, regionális problémát is jelent. Így – az FCS-hez hasonlóan – a betegek és a hozzátartozók figyelmének felhívása erre, az önmagában is komoly egészségügyi kockázattal járó állapotról, kiemelt jelentőséggel kell, hogy bírjon [75].

Az FH páciensek Lp(a) értékeit viszonylag kevés tanulmányban vizsgálták részletesen, ugyanakkor az eddig ismert adatok a mi eredményeinkkel jelentős összhangban vannak [70,76,77]. Megemlítendő az is, hogy az irodalmi adatokkal szintén egyezően, a nők körében szignifikánsan magasabb Lp(a) értékeket találtunk [76]. Ez a különbség egyébként nem csak az FH betegek körében figyelhető meg, de a CVD [78] és PAD [79] betegeknél, valamint az általános populációban is [80]. Bár a nemek közötti különbség pontos oka nem ismert, de feltételezhető, hogy az ösztrogén apo(a) fehérje expressziójára kifejtett hatása felel az eltérésért [81].

Bár az Lp(a) mérések alacsony száma gátolta a szignifikáns eltérések azonosítását az FH betegeknél, az eltérések iránya alapján valószínűsíthető, hogy az extrém magas Lp(a) értékek ezen betegek esetén is tovább növelik a kardiovaszkuláris események kockázatát. Ezen megállapítások megerősítéséhez, így mind az irány megerősítéséhez, mind pedig a pontos kockázatbecsléséhez azonban további elemzések szükségesek. Az utóbbi években számos kockázatbecslő módszert fejlesztettek ki az FH betegek kardiovaszkuláris rizikójának meghatározására. Ilyen például a Spanyolországban és Franciaországban elfogadott SAFEHEART rizikóbecslő [82,83], vagy a MONTREAL FH SCORE, melyet kanadai FH

betegekre validáltak [84]. Ahhoz, hogy hasonló kockázatbecslés a magyar populációra is alkalmazható legyen, elengedhetetlen az egyes kockázati tényezők előfordulási arányának meghatározása. A mi kutatásunk az első, amely tömeges adatokon számszerűsített előfordulási arányokat és kockázatok határozott meg hazánkban, mely – más régiók adatait is felhasználva – kiindulásul szolgálhat egy általános összefüggésrendszer, illetve rizikóbecslés létrehozásához.

Adataink alapján megállapítható, hogy az érrendszeri megbetegedések igen magas előfordulása az észak-alföldi régió FH populációján belül nagyrészt a legfőbb kardiovaszkuláris kockázati tényezők (úm. a dohányzás, a magas vérnyomás és az elhízás) gyakori, mintegy klaszter-szerű előfordulásával magyarázható. Azon betegek, akik többszörös metabolikus kockázati tényezőkkel rendelkeznek, különösen hasznát láthatják az új kezelési lehetőségeknek és a személyre szabott terápiás stratégiának [85]. A PCSK9 gátlók, mind monoklonális antitest, mind rövid interferáló RNS formában körülbelül 26%-kal képesek csökkenteni az Lp(a) szintet, mely azonban az extrém magas Lp(a) értékekkel rendelkező páciensek esetén nem elegendő [86]. A közeljövő gyógyszerei, így az antiszenz oligonukleotidok és újabb rövid interferáló RNS készítmények specifikusan, más tényezőre nem hatva, akár 80-90%-os Lp(a) csökkentést is elérhetnek [87]. A jelenleg is folyó klinikai kutatások függvényében ezek az új hatóanyagok jelentős segítséget nyújthatnak a magas Lp(a) szérumszinttel rendelkező FH és nem FH betegeknek is. Ezen új terápiás lehetőségek költsége azonban igen magas, így az egyéb, módosítható kockázati tényezők és káros szokások megfelelő kontrollja kiemelkedő jelentőségű. Kétségtelenül azok a nemdohányzó, optimális testsúllyal és LDL-C értékekkel rendelkező betegek, akik hipertóniája és diabétesze is egyensúlyban van, sokkal nagyobb mértékben profitálhatnak az új kezelési lehetőségekből, melyet az általunk végzett HR kalkulációk is szorosan alátámasztanak. Így az egyre hatékonyabb terápiás megoldások mellett a hagyományos kezelési módok, úm. a másodlagos kockázati tényezők kizárása, az egészséges életmód, diéta, a rendszeres fizikai aktivitással egybekötött testsúlykontroll, valamint az egyéb kardiovaszkuláris rizikófaktorok együttes kezelése továbbra is sarokkövei az FH betegek terápiájának és a kardiovaszkuláris prevenciónak [16].

A kutatás korlátai

A nagy adattömegekre épülő retrospektív kutatások alapvető hiányossága, hogy magára az adatminőségre és az adat strukturális jellemzőire nem tudunk hatással lenni. Így

vizsgálataink során sem volt elérhető számos olyan adat, amelyekre a pontos score számításokhoz ugyanakkor szükségünk lett volna. Így pl. a családi anamnézis csak akkor volt hozzáférhető, amennyiben az eredeti adatfelvétel során ez rögzítésre került, mely azonban csak az esetek elhanyagolható részére volt igaz. Szintén nem volt hozzáférésünk a betegek esetlegesen korábban készült genetikai elemzéséhez. Ugyanakkor, még nem publikált vizsgálatainkban részletes genetikai vizsgálatokat végeztünk, melyek magas találati aránnyal jelezték az adatbányászat és gépi tanulás során megtalált pácienseket. Az ismert és igazolt patogén mutációk felhasználása pontosabb prevalenciát és a betegeket jobban felismerő matematikai modelleket eredményezhet.

Az FH betegek pontos számának becsléséhez, a DLCN pontérték kiszámításához fontos az LDL-C szintek pontos ismerete. Amennyiben a beteg lipidcsökkentő kezelésben részesül, az LDL-C szint módosítása javasolt a DLCN pontszámok meghatározása során [88], ugyanakkor arról nem volt információnk, hogy a háziorvosok kezdeményezték-e ilyen terápiát a beteg beutalása előtt. Emiatt a kezelés előtt mért lipidszintek átlagával számoltunk minden ilyen esetben. Megemlítendő, hogy egyes külföldi tanulmányok szerint a háziorvosok képesek pontosan azonosítani az FH betegeket a DLCN pontszámok alapján [86, 87], de a magyar egészségügyi rendszer szabályozása, valamint a finanszírozási problémák és az LDL-C mérések területi hiányosságai miatt a háziorvosok legtöbbször korán beküldik a hiperlipidémiás betegeket a szakrendelésekre, amely a hibás kalkulációk számát ha nem is szünteti meg, de minimalizálja.

Problémát jelenthet az is, hogy a vizsgált klinikai beteganyag és a valós társadalmi populáció között mennyire szoros és/vagy arányos a kapcsolat, azaz a klinikai beteganyag mennyire képezi le a társadalmi megoszlást. Fentebb már említettem, hogy a regionális populáció közel 80%-a megfordult a vizsgált intézményekben, így itt ez a hiba feltehetőleg nem okozott jelentős problémát – azonban az adat tulajdonságai miatt eredményeink inkább az idősebb korosztályokat képezik le, mivel a fiatalabb korosztály ritkábban jár kórházba, emellett a laboratóriumi vizsgálataik is kevésbé széleskörűek és részletes anamnézis is ritkábban kerül felvételre. További eltérést okozhat, hogy bizonyos vizsgálatok (például az Lp(a) meghatározása) a betegeknél nem voltak általánosan elérhetőek, hanem csak bizonyos, legtöbbször idősebb pácienseknél ellenőrizték, akiknél felmerült a kardiovaszkuláris betegség gyanúja, vagy igazolt CVD betegek voltak.

Tárgyszavak

Tárgyszavak: ritka anyagcsere betegségek, familiáris hiperkoleszterinémia, familiáris kilomikronémia szindróma, lipoprotein(a), gépi tanulás, adatbányászat, számítógéppel támogatott diagnosztika

Keywords: rare metabolic disorders, familial hypercholesterolemia, familial chylomicronemia syndrome, lipoprotein(a), machine learning, data mining, computer-aided diagnostics

Összefoglalás

Az adatbányászat és gépi tanulás módszerei nagy klinikai és kórházi adattömegben alkalmazva, rendkívül hasznosnak bizonyultak mind a ritka betegségek prevalenciájának becslésében, mind pedig ezen ritka kórképeket hordozó betegek azonosításában.

Prevalencia becsléseink alapján az FH és az FCS is gyakoribbnak bizonyult a korábbi irodalmi adatokhoz képest, ugyanakkor újabb vizsgálatok alapján feltételezhető, hogy a magasabb prevalencia lehet a valós. Az ilyen típusú elemzés során maguknak a pontrendszernek a teljesítménye is vizsgálható és sikerült igazolnunk, hogy mind a DCLNS, mind pedig a Moulin-féle FCS score nagyon hatékony az FH, illetve az FCS betegek diagnosztizálásában, a súlyos genetikai érintettségek által okozott kórképek identifikálásában. Utóbbi pontrendszer az FCS és MFCS esetek szétválasztásában is megbízhatóan használható.

Az FCS score mélyebb vizsgálatánál arra jutottunk, hogy a gépi tanulás eszközszerével maga a pontrendszer is tovább fejleszthető. Mindez azért is tűnik hasznosnak, mivel a pontszám kiszámítása során olyan információkra is támaszkodnunk kell, melyek gyakran az egészségügyi adatbázisokban sajnálatosan nem mindig elérhetőek. A nagy adattömegben történő tanítás a betegség kórlefolyásának olyan összefüggéseit is feltárhatja, amelyek korábban nem voltak ismertek, vagy nem volt komolyabb jelentőség tulajdonítva nekik. Munkánk alapján – véleményünk szerint fontos üzenetként – a szérum összkoleszterin szintjének ismerete az FCS pontosabb diagnózisát és a betegek biztosabb azonosítását teheti lehetővé.

A módszer alkalmazhatóságát részben nehezíti az egyes gépi tanulási modellek “black box” jellege, azaz, hogy a pontos döntési utak a modell komplex súlyozása és számos paramétere miatt nehezen vagy egyáltalán nem fejthetőek vissza. Mégis ezek a modern megoldások jelenthetik a jövő tömeges, de személyre szabott diagnosztikájának fontos elemét leginkább azért, mert ezek alapvető tulajdonsága a gyors és tömeges alkalmazhatóság. A legtöbb beteg általában többször is megfordul valamelyik egészségügyi szolgáltatónál, mely során standard orvosi vizsgálat és rutin vérvétel általában történik. Ezek az adatok pedig legtöbbször elégségesek arra, hogy gyors becslés alapján az FH vagy az FCS kockázata meghatározható legyen (általában ennél jóval több adat is rendelkezésre áll, mire a beteg tényleges diagnózisát megállapítják). Ez lehetőséget ad más ritka betegségek korábbi azonosítására, ezáltal a korai kezelés megkezdésére, a potenciálisan súlyos kardiovaszkuláris

szövődmények elkerülésére, hozzájárulva ezzel a társadalom tagjainak egészségben töltött éveinek növekedéséhez.

A korai, számítógéppel támogatott vizsgálatok szintén segíthetik a megfelelő jelöltek kiválasztását további diagnosztikai és genetikai vizsgálatokra, mely által a diagnosztikára fordított erőforrások csökkenhetnek. Kutatásaink során igazoltuk, hogy az adatbányászat és a gépi tanulás módszerei hatékonyak a kockázati tényezők számszerűsítésében is, így a diagnosztika mellett a tudományos kutatás fontos elemét is jelenthetik. Az egészségügyi szolgáltató intézmények hatékony adatstruktúrák (adattárházak és adattavak) és megfelelő, támogató ML rendszerek kialakításával mind a diagnosztika, mind pedig a tudományos kutatás során is jelentős előnyre tehet szert az ilyen megoldásokat nem alkalmazó intézményekkel szemben.

Summary

Applied to large amounts of hospital data, methods of data mining and machine learning have proven to be extremely useful both in estimating the prevalence of rare diseases and identifying the patients.

During our prevalence estimation process, we found that both FH and FCS are more common than they were suggested by earlier literature data, however, recent studies indicate higher prevalence. During this type of analysis, the performance of the score systems themselves can also be assessed numerically. Indeed, we confirmed that both DCLN and FCS scoring are very effective in the clinical practice in diagnosing FH and FCS patients, separating FCS and MFCS cases and, finding serious genetic involvements.

We also came to the conclusion that the FCS score system can be further improved with the tools of machine learning. Although score systems are more than just fast numerical scoring, much information is not available in hospital databases including family history or detailed genetic data, therefore improvement of scoring efficacy is of major interest. Training on big data may reveal previously unknown relationships in disease appearance or course. Indeed, we now recommend a more detailed consideration of the serum cholesterol levels when establishing the diagnosis of FCS.

The applicability of the method is partly hampered by the "black box" nature of the individual ML models (that is, the exact decision paths are difficult or impossible to decipher due to the complex parameters of the model). Yet these modern solutions may represent important elements of the mass-personalized diagnostics in the future, mostly because they are rapid and easy to apply. Additionally, most patients turn up at health care providers several times participating in standard medical examinations and routine blood sampling. These data are usually sufficient enough to properly determine the risk of FH or FCS based on a quick risk assessment, however, in most cases, much more data is available by the time the actual diagnosis is established. Risk assessment may also further improve the screening of these and other rare diseases, initiate early treatment and help to avoid serious and life-threatening cardiovascular complications, as well as to increase the life expectancy.

Early computer-assisted screening tests may also increase the selection of appropriate candidates for further diagnostic and genetic testing, thereby reducing resources devoted to individual diagnostic efforts. During our research, we also managed to prove that data mining and machine learning are also effective for quantifying exact risk factors, thus adding important elements of scientific research to the diagnostic process.. By the development of

effective data structures including data warehouses and data lakes and suitable, supporting ML systems, health care providers may gain an irreplaceable advantage over the ones without such solutions, both during diagnostics and scientific research.

Felhasznált irodalom

1. Mackenbach JP, Kulhánová I, Menvielle G, Bopp M, Borrell C, Costa G, Deboosere P, Esnaola S, Kalediene R, Kovacs K, Leinsalu M, Martikainen P, Regidor E, Rodriguez-Sanz M, Strand BH, Hoffmann R, Eikemo TA, Östergren O, Lundberg O. Trends in inequalities in premature mortality: a study of 3.2 million deaths in 13 European countries. *J Epidemiol Community Health*. 2015 Mar 1;69(3):207–17.
2. Blakely T, Disney G, Atkinson J, Teng A, Mackenbach JP. A Typology for Charting Socioeconomic Mortality Gradients: “Go Southwest.” *Epidemiol Camb Mass*. 2017 Jul;28(4):594–603.
3. Cholesterol Treatment Trialists’ (CTT) Collaboration, Fulcher J, O’Connell R, Voysey M, Emberson J, Blackwell L, Mihaylova B, Simes J, Collins R, Kirby A, Colhoun H, Braunwald E, La Rosa J, Pedersen TR, Tonkin A, Davis B, Sleight P, Franzosi MG, Baigent C, Keech A. Efficacy and safety of LDL-lowering therapy among men and women: meta-analysis of individual data from 174,000 participants in 27 randomised trials. *Lancet Lond Engl*. 2015 Apr 11;385(9976):1397–405.
4. Hartgers ML, Ray KK, Hovingh GK. New Approaches in Detection and Treatment of Familial Hypercholesterolemia. *Curr Cardiol Rep*. 2015 Dec;17(12):109.
5. Goldstein JL, Schrott HG, Hazzard WR, Bierman EL, Motulsky AG. Hyperlipidemia in coronary heart disease. II. Genetic analysis of lipid levels in 176 families and delineation of a new inherited disorder, combined hyperlipidemia. *J Clin Invest*. 1973 Jul;52(7):1544–68.
6. Sjouke B, Kusters DM, Kindt I, Besseling J, Defesche JC, Sijbrands EJG, Roeters van Lennep JE, Stalenhoef AFH, Wiegman A, de Graaf J, Fouchier SW, Kastelein JJP, Hovingh GK. Homozygous autosomal dominant hypercholesterolaemia in the Netherlands: prevalence, genotype-phenotype relationship, and clinical outcome. *Eur Heart J*. 2015 Mar 1;36(9):560–5.
7. Akioyamen LE, Genest J, Shan SD, Reel RL, Albaum JM, Chu A, Tu JV. Estimating the prevalence of heterozygous familial hypercholesterolaemia: a systematic review and meta-analysis. *BMJ Open*. 2017 Sep 1;7(9):e016461.
8. Mundal L, Sarancic M, Ose L, Iversen PO, Borgan JK, Veierød MB, Leren TP, Retterstøl K. Mortality among patients with familial hypercholesterolemia: a registry-based study in Norway, 1992-2010. *J Am Heart Assoc*. 2014 Dec

2;3(6):e001236.

9. Goldberg RB, Chait A. A Comprehensive Update on the Chylomicronemia Syndrome. *Front Endocrinol* [Internet]. 2020 [cited 2023 Aug 28];11. Available from: <https://www.frontiersin.org/articles/10.3389/fendo.2020.593931>
10. Hegele RA, Berberich AJ, Ban MR, Wang J, Digenio A, Alexander VJ, D'Erasmus L, Arca M, Jones A, Bruckert E, Stroes ES, Bergeron J, Civeira F, Witztum JL, Gaudet D. Clinical and biochemical features of different molecular etiologies of familial chylomicronemia. *J Clin Lipidol*. 2018;12(4):920-927.e4.
11. Beigneux AP, Miyashita K, Ploug M, Blom DJ, Ai M, Linton MF, Khovidhunkit W, Dufour R, Garg A, McMahon MA, Pullinger CR, Sandoval NP, Hu X, Allan CM, Larsson M, Machida T, Murakami M, Reue K, Tontonoz P, Goldberg IJ, Moulin P, Charrière S, Fong LG, Nakajima K, Young SG. Autoantibodies against GPIHBP1 as a Cause of Hypertriglyceridemia. *N Engl J Med*. 2017 Apr 27;376(17):1647–58.
12. Dionisi-Vici C, Shteyer E, Niceta M, Rizzo C, Pode-Shakked B, Chillemi G, Bruselles A, Semeraro M, Barel O, Eyal E, Kol N, Haberman Y, Lahad A, Diomedi-Camassei F, Marek-Yagel D, Rechavi G, Tartaglia M, Anikster Y. Expanding the molecular diversity and phenotypic spectrum of glycerol 3-phosphate dehydrogenase 1 deficiency. *J Inherit Metab Dis*. 2016 Sep;39(5):689–95.
13. Hegele RA, Ginsberg HN, Chapman MJ, Nordestgaard BG, Kuivenhoven JA, Aversa M, Borén J, Bruckert E, Catapano AL, Descamps OS, Hovingh GK, Humphries SE, Kovanen PT, Masana L, Pajukanta P, Parhofer KG, Raal FJ, Ray KK, Santos RD, Stalenhoef AFH, Stroes E, Taskinen MR, Tybjærg-Hansen A, Watts GF, Wiklund O, European Atherosclerosis Society Consensus Panel. The polygenic nature of hypertriglyceridaemia: implications for definition, diagnosis, and management. *Lancet Diabetes Endocrinol*. 2014 Aug;2(8):655–66.
14. Moulin P, Dufour R, Aversa M, Arca M, Cefalù AB, Noto D, D'Erasmus L, Di Costanzo A, Marçais C, Alvarez-Sala Walther LA, Banach M, Borén J, Cramb R, Gouni-Berthold I, Hughes E, Johnson C, Pintó X, Reiner Ž, van Lennep JR, Soran H, Stefanutti C, Stroes E, Bruckert E. Identification and diagnosis of patients with familial chylomicronaemia syndrome (FCS): Expert panel recommendations and proposal of an “FCS score.” *Atherosclerosis*. 2018 Aug;275:265–72.
15. Paquette M, Bernard S, Hegele RA, Baass A. Chylomicronemia: Differences between familial chylomicronemia syndrome and multifactorial chylomicronemia. *Atherosclerosis*. 2019 Apr;283:137–42.

16. Berberich AJ, Hegele RA. A Modern Approach to Dyslipidemia. *Endocr Rev.* 2022 Jul 13;43(4):611–53.
17. Beaumont JL, Carlson LA, Cooper GR, Fejfar Z, Fredrickson DS, Strasser T. Classification of hyperlipidaemias and hyperlipoproteinaemias. *Bull World Health Organ.* 1970;43(6):891–915.
18. Aygun S, Tokgozoglul L. Comparison of Current International Guidelines for the Management of Dyslipidemia. *J Clin Med.* 2022 Dec 6;11(23):7249.
19. Vaezi Z, Amini A. Familial Hypercholesterolemia. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 [cited 2023 Aug 28]. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK556009/>
20. Guo Q, Feng X, Zhou Y. PCSK9 Variants in Familial Hypercholesterolemia: A Comprehensive Synopsis. *Front Genet.* 2020 Sep 23;11:1020.
21. Austin MA, Hutter CM, Zimmern RL, Humphries SE. Genetic causes of monogenic heterozygous familial hypercholesterolemia: a HuGE prevalence review. *Am J Epidemiol.* 2004 Sep 1;160(5):407–20.
22. Paragh G, Németh Á, Harangi M, Banach M, Fülöp P. Causes, clinical findings and therapeutic options in chylomicronemia syndrome, a special form of hypertriglyceridemia. *Lipids Health Dis.* 2022 Feb 10;21(1):21.
23. Mead JR, Irvine SA, Ramji DP. Lipoprotein lipase: structure, function, regulation, and role in disease. *J Mol Med Berl Ger.* 2002 Dec;80(12):753–69.
24. Young SG, Davies BSJ, Voss CV, Gin P, Weinstein MM, Tontonoz P, Reue K, Bensadoun A, Fong LG, Beigneux AP. GPIHBP1, an endothelial cell transporter for lipoprotein lipase. *J Lipid Res.* 2011 Nov;52(11):1869–84.
25. Calandra S, Priore Oliva C, Tarugi P, Bertolini S. APOA5 and triglyceride metabolism, lesson from human APOA5 deficiency. *Curr Opin Lipidol.* 2006 Apr;17(2):122–7.
26. Huff MW, Hegele RA. Apolipoprotein C-III: going back to the future for a lipid drug target. *Circ Res.* 2013 May 24;112(11):1405–8.
27. Mattijssen F, Kersten S. Regulation of triglyceride metabolism by Angiopoietin-like proteins. *Biochim Biophys Acta.* 2012 May;1821(5):782–9.
28. Sandhu S, Al-Sarraf A, Taraboanta C, Frohlich J, Francis GA. Incidence of pancreatitis, secondary causes, and treatment of patients referred to a specialty lipid clinic with severe hypertriglyceridemia: a retrospective cohort study. *Lipids Health Dis.* 2011 Sep 11;10:157.
29. Tan PN. Introduction to Data Mining. Pearson; 800 p.

30. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge University Press; 2017. book.
31. Bird S, Klein E, Loper E. Natural Language Processing with Python. 1st ed. O'Reilly Media, Inc.; 2009. 512 p.
32. Friedl J. Mastering Regular Expressions. O'Reilly Media, Inc.; 2006.
33. Eisenstein J. Introduction to Natural Language Processing. MIT Press; 2019. 535 p.
34. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2013 [cited 2023 Aug 28]. Available from: https://papers.nips.cc/paper_files/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html
35. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is All you Need. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2017 [cited 2023 Aug 28]. Available from: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
36. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In 2019 [cited 2023 Aug 28]. p. 4171–86. Available from: <https://aclanthology.org/N19-1423>
37. Garipov T, Izmailov P, Podoprikin D, Vetrov DP, Wilson AG. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2018. Available from: https://proceedings.neurips.cc/paper_files/paper/2018/file/be3087e74e9100d4bc4c6268cdbe8456-Paper.pdf
38. Choromanska A, Henaff Mi, Mathieu M, Ben Arous G, LeCun Y. The Loss Surfaces of Multilayer Networks. In: Lebanon G, Vishwanathan SVN, editors. Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics [Internet]. San Diego, California, USA: PMLR; 2015. p. 192–204. (Proceedings of Machine Learning Research; vol. 38). Available from: <https://proceedings.mlr.press/v38/choromanska15.html>
39. Heckerman D. A Tutorial on Learning with Bayesian Networks. In: Holmes DE, Jain LC, editors. Innovations in Bayesian Networks: Theory and Applications [Internet].

- Berlin, Heidelberg: Springer; 2008 [cited 2023 Sep 11]. p. 33–82. (Studies in Computational Intelligence). Available from: https://doi.org/10.1007/978-3-540-85066-3_3
40. Johns BT, Jamieson RK. A Large-Scale Analysis of Variance in Written Language. *Cogn Sci*. 2018;42(4):1360–74.
 41. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armañanzas R, Santafé G, Pérez A, Robles V. Machine learning in bioinformatics. *Brief Bioinform*. 2006 Mar;7(1):86–112.
 42. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling [Internet]. arXiv; 2014 [cited 2023 Nov 4]. Available from: <http://arxiv.org/abs/1412.3555>
 43. Németh Á, Daróczy B, Juhász L, Fülöp P, Harangi M, Paragh G. Assessment of Associations Between Serum Lipoprotein (a) Levels and Atherosclerotic Vascular Diseases in Hungarian Patients With Familial Hypercholesterolemia Using Data Mining and Machine Learning. *Front Genet*. 2022;13:849197.
 44. Vapnik VN, Chervonenkis AYa. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory Probab Its Appl*. 1971 Jan;16(2):264–80.
 45. Nagarajan V, Kolter JZ. Uniform convergence may be unable to explain generalization in deep learning. In: *Advances in Neural Information Processing Systems* [Internet]. Curran Associates, Inc.; 2019 [cited 2023 Sep 1]. Available from: https://proceedings.neurips.cc/paper_files/paper/2019/hash/05e97c207235d63ceb1db43c60db7bbb-Abstract.html
 46. Devroye L, Györfi L, Lugosi G. A Probabilistic Theory of Pattern Recognition [Internet]. New York, NY: Springer; 1996 [cited 2023 Sep 1]. (Karatzas I, Yor M, editors. *Stochastic Modelling and Applied Probability*; vol. 31). Available from: <http://link.springer.com/10.1007/978-1-4612-0711-5>
 47. Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J Comput Syst Sci*. 1997 Aug 1;55(1):119–39.
 48. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. *Ann Stat*. 2001;29(5):1189–232.
 49. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995 Sep 1;20(3):273–97.
 50. Montúfar G, Pascanu R, Cho K, Bengio Y. On the Number of Linear Regions of Deep Neural Networks [Internet]. arXiv; 2014 [cited 2023 Sep 1]. Available from: <http://arxiv.org/abs/1402.1869>

51. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization [Internet]. arXiv; 2017 [cited 2023 Sep 1]. Available from: <http://arxiv.org/abs/1412.6980>
52. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: Association for Computing Machinery; 2016 [cited 2023 Sep 1]. p. 785–94. (KDD '16). Available from: <https://dl.acm.org/doi/10.1145/2939672.2939785>
53. Damgaard D, Larsen ML, Nissen PH, Jensen JM, Jensen HK, Soerensen VR, Jensen LG, Faergeman O. The relationship of molecular genetic to clinical diagnosis of familial hypercholesterolemia in a Danish population. *Atherosclerosis*. 2005 May;180(1):155–60.
54. Brahm AJ, Hegele RA. Chylomicronaemia--current diagnosis and future therapies. *Nat Rev Endocrinol*. 2015 Jun;11(6):352–62.
55. Bell DA, Watts GF. Progress in the care of familial hypercholesterolaemia: 2016. *Med J Aust*. 2016;205(5):232–6.
56. Arca M. Old challenges and new opportunities in the clinical management of heterozygous familial hypercholesterolemia (HeFH): The promises of PCSK9 inhibitors. *Atherosclerosis*. 2017 Jan;256:134–45.
57. Pallazola VA, Sajja A, Derenbecker R, Ogunmoroti O, Park J, Sathiyakumar V, Martin SS. Prevalence of familial chylomicronemia syndrome in a quaternary care center. *Eur J Prev Cardiol*. 2020 Dec;27(19):2276–8.
58. Patni N, Li X, Adams-Huet B, Garg A. The prevalence and etiology of extreme hypertriglyceridemia in children: Data from a tertiary children's hospital. *J Clin Lipidol*. 2018;12(2):305–10.
59. Ohm F, Vogel D, Sehner S, Wijnen-Meijer M, Harendza S. Details acquired from medical history and patients' experience of empathy--two sides of the same coin. *BMC Med Educ*. 2013 May 9;13:67.
60. Paragh G, Harangi M, Karányi Z, Daróczy B, Németh Á, Fülöp P. Identifying patients with familial hypercholesterolemia using data mining methods in the Northern Great Plain region of Hungary. *Atherosclerosis*. 2018 Oct;277:262–6.
61. Dallinga-Thie GM, Kroon J, Borén J, Chapman MJ. Triglyceride-Rich Lipoproteins and Remnants: Targets for Therapy? *Curr Cardiol Rep*. 2016;18:67.
62. Maltais M, Brisson D, Gaudet D. Non-Alcoholic Fatty Liver in Patients with Chylomicronemia. *J Clin Med*. 2021 Feb 9;10(4):669.

63. Neefjes LA, Ten Kate GJR, Alexia R, Nieman K, Galema-Boers AJ, Langendonk JG, Weustink AC, Mollet NR, Sijbrands EJ, Krestin GP, de Feyter PJ. Accelerated subclinical coronary atherosclerosis in patients with familial hypercholesterolemia. *Atherosclerosis*. 2011 Dec;219(2):721–7.
64. Silva PRS, Jannes CE, Marsiglia JDC, Krieger JE, Santos RD, Pereira AC. Predictors of cardiovascular events after one year of molecular screening for Familial hypercholesterolemia. *Atherosclerosis*. 2016 Jul;250:144–50.
65. Tatò F, Keller C, Schuster H, Spengel F, Wolfram G, Zöllner N. Relation of lipoprotein(a) to coronary heart disease and duplexsonographic findings of the carotid arteries in heterozygous familial hypercholesterolemia. *Atherosclerosis*. 1993 Jun;101(1):69–77.
66. Dyrbuś K, Gąsior M, Desperak P, Osadnik T, Nowak J, Banach M. The prevalence and management of familial hypercholesterolemia in patients with acute coronary syndrome in the Polish tertiary centre: Results from the TERCET registry with 19,781 individuals. *Atherosclerosis*. 2019 Sep;288:33–41.
67. Korneva V, Kuznetsova T, Julius U. Efficiency and problems of statin therapy in patients with heterozygous familial hypercholesterolemia. *Atheroscler Suppl*. 2019 Dec;40:79–87.
68. Vlad CE, Foia L, Florea L, Costache II, Covic A, Popescu R, Reurean-Pintilei D, Covic A. Evaluation of cardiovascular risk factors in patients with familial hypercholesterolemia from the North-Eastern area of Romania. *Lipids Health Dis*. 2021 Jan 11;20:4.
69. Bertolini S, Pisciotta L, Rabacchi C, Cefalù AB, Noto D, Fasano T, Signori A, Fresa R, Averna M, Calandra S. Spectrum of mutations and phenotypic expression in patients with autosomal dominant hypercholesterolemia identified in Italy. *Atherosclerosis*. 2013 Apr;227(2):342–8.
70. Mehta R, Martagon AJ, Galan Ramirez GA, Antonio-Villa NE, Vargas-Vázquez A, Elias-Lopez D, Gonzalez-Retana G, Rodríguez-Encinas B, Ceballos-Macías JJ, Romero-Zazueta A, Martinez-Alvarado R, Morales-Portano JD, Alvarez-Lopez H, Sauque-Reyna L, Gomez-Herrera LG, Simental-Mendia LE, Garcia-Aguilar H, Ramirez-Cooremans E, Peña-Aparicio B, Mendoza-Zubieta V, Carrillo-Gonzalez PA, Ferreira-Hermosillo A, Caracas-Portilla N, Jimenez-Dominguez G, Ruiz-Garcia AY, Arriaga-Cazares HE, Gonzalez-Gonzalez JR, Mendez-Valencia CV, Padilla FG, Madriz-Prado R, De Los Rios-Ibarra MO, Vazquez-Cardenas A, Arjona-Villicaña RD,

- Acevedo-Rivera KJ, Allende-Carrera R, Alvarez JA, Amezcua-Martinez JC, de Los Reyes Barrera-Bustillo M, Carazo-Vargas G, Contreras-Chacon R, Figueroa-Andrade MH, Flores-Ortega A, Garcia-Alcala H, Garcia de Leon LE, Garcia-Guzman B, Garduño-Garcia JJ, Garnica-Cuellar JC, Gomez-Cruz JR, Hernandez-Garcia A, Holguin-Almada JR, Juarez-Herrera U, Lugo-Sobrevilla F, Marquez-Rodriguez E, Martinez-Sibaja C, Medrano-Rodriguez AB, Morales-Oyervides JC, Perez-Vazquez DI, Reyes-Rodriguez EA, Robles-Osorio ML, Rosas-Saucedo J, Torres-Tamayo M, Valdez-Talavera LA, Vera-Arroyo LE, Zepeda-Carrillo EA, Aguilar-Salinas CA, Mexican Familial Hypercholesterolemia Group. Familial hypercholesterolemia in Mexico: Initial insights from the national registry. *J Clin Lipidol*. 2021;15(1):124–33.
71. Besseling J, Kindt I, Hof M, Kastelein JJP, Hutten BA, Hovingh GK. Severe heterozygous familial hypercholesterolemia and risk for cardiovascular disease: a study of a cohort of 14,000 mutation carriers. *Atherosclerosis*. 2014 Mar;233(1):219–23.
 72. Kayikcioglu M, Tokgozoglu L, Dogan V, Ceyhan C, Tuncel A, Kutlu M, Onrat E, Alici G, Akbulut M, Celik A, Yesilbursa D, Sahin T, Sonmez A, Ozdogan O, Temizhan A, Kilic S, Bayram F, Sabuncu T, Coskun FY, Ildizli M, Durakoglugil E, Kirilmaz B, Yilmaz MB, Yigit Z, Yildirim AB, Gedikli O, Topcu S, Oğuz A, Demir M, Yenerçag M, Yıldırım A, Demircan S, Yilmaz M, Kaynar LG, Aktan M, Durmus RB, Gokce C, Ozcebe Oİ, Akyol TK, Okutan H, Sag S, Gul OO, Salcioglu Z, Altunkeser BB, Kuku I, Yasar HY, Kurtoglu E, Kose MD, Demircioglu S, Pekkolay Z, Ilhan O, Can LH. What have we learned from Turkish familial hypercholesterolemia registries (A-HIT1 and A-HIT2)? *Atherosclerosis*. 2018 Oct;277:341–6.
 73. Vohnout B, Fábryová Ľ, Klabník A, Kadurová M, Bálinth K, Kozárová M, Bugáňová I, Sirotiaková J, Rašlová K. Treatment pattern of familial hypercholesterolemia in Slovakia: Targets, treatment and obstacles in common practice. *Atherosclerosis*. 2018 Oct;277:323–6.
 74. Besseling J, Kastelein JJP, Defesche JC, Hutten BA, Hovingh GK. Association between familial hypercholesterolemia and prevalence of type 2 diabetes mellitus. *JAMA*. 2015 Mar 10;313(10):1029–36.
 75. Ceska R, Latkovskis G, Ezhov MV, Freiburger T, Lalic K, Mitchenko O, Paragh G, Petrulioniene Z, Pojskic B, Raslova K, Shek AB, Vohnout B, Altschmiedova T, Todorovova V. The Impact of the International Cooperation On Familial Hypercholesterolemia Screening and Treatment: Results from the ScreenPro FH Project. *Curr Atheroscler Rep*. 2019 Jun 22;21(9):36.

76. Lingenhel A, Kraft HG, Kotze M, Peeters AV, Kronenberg F, Kruse R, Utermann G. Concentrations of the atherogenic Lp(a) are elevated in FH. *Eur J Hum Genet EJHG*. 1998 Jan;6(1):50–60.
77. Alonso R, Andres E, Mata N, Fuentes-Jiménez F, Badimón L, López-Miranda J, Padró T, Muñiz O, Díaz-Díaz JL, Mauri M, Ordovás JM, Mata P, SAFEHEART Investigators. Lipoprotein(a) levels in familial hypercholesterolemia: an important predictor of cardiovascular disease independent of the type of LDL receptor mutation. *J Am Coll Cardiol*. 2014 May 20;63(19):1982–9.
78. Virani SS, Brautbar A, Davis BC, Nambi V, Hoogeveen RC, Sharrett AR, Coresh J, Mosley TH, Morrisett JD, Catellier DJ, Folsom AR, Boerwinkle E, Ballantyne CM. Associations Between Lipoprotein(a) Levels and Cardiovascular Outcomes in African Americans and Caucasians: The Atherosclerosis Risk in Communities (ARIC) Study. *Circulation*. 2012 Jan 17;125(2):241–9.
79. Forbang NI, Criqui MH, Allison MA, Ix JH, Steffen BT, Cushman M, Tsai MY. Sex and ethnic differences in the associations between lipoprotein(a) and peripheral arterial disease in the Multi-Ethnic Study of Atherosclerosis. *J Vasc Surg*. 2016 Feb;63(2):453–8.
80. Banerjee D, Wong EC, Shin J, Fortmann SP, Palaniappan L. Racial and Ethnic Variation in Lipoprotein (a) Levels among Asian Indian and Chinese Patients. *J Lipids*. 2011 May 23;2011:e291954.
81. Boffelli D, Zajchowski DA, Yang Z, Lawn RM. Estrogen modulation of apolipoprotein(a) expression. Identification of a regulatory element. *J Biol Chem*. 1999 May 28;274(22):15569–74.
82. Pérez de Isla L, Alonso R, Mata N, Fernández-Pérez C, Muñiz O, Díaz-Díaz JL, Saltijeral A, Fuentes-Jiménez F, de Andrés R, Zambón D, Piedecausa M, Cepeda JM, Mauri M, Galiana J, Brea Á, Sanchez Muñoz-Torrero JF, Padró T, Argueso R, Miramontes-González JP, Badimón L, Santos RD, Watts GF, Mata P. Predicting Cardiovascular Events in Familial Hypercholesterolemia: The SAFEHEART Registry (Spanish Familial Hypercholesterolemia Cohort Study). *Circulation*. 2017 May 30;135(22):2133–44.
83. Gallo A, Charriere S, Vimont A, Chapman MJ, Angoulvant D, Boccara F, Cariou B, Carreau V, Carrié A, Bruckert E, Béliard S, French REgistry of Familial hyperCHOLesterolemia (REFERCHOL) investigators. SAFEHEART risk-equation and cholesterol-year-score are powerful predictors of cardiovascular events in French

- patients with familial hypercholesterolemia. *Atherosclerosis*. 2020 Aug;306:41–9.
84. Paquette M, Dufour R, Baass A. The Montreal-FH-SCORE: A new score to predict cardiovascular events in familial hypercholesterolemia. *J Clin Lipidol*. 2017;11(1):80–6.
 85. Giglio RV, Stoian AP, Haluzik M, Pafili K, Patti AM, Rizvi AA, Ciaccio M, Papanas N, Rizzo M. Novel molecular markers of cardiovascular disease risk in type 2 diabetes mellitus. *Biochim Biophys Acta Mol Basis Dis*. 2021 Aug 1;1867(8):166148.
 86. Spolitu S, Dai W, Zadroga JA, Ozcan L. Proprotein convertase subtilisin/kexin type 9 and lipid metabolism. *Curr Opin Lipidol*. 2019 Jun;30(3):186–91.
 87. Viney NJ, van Capelleveen JC, Geary RS, Xia S, Tami JA, Yu RZ, Marcovina SM, Hughes SG, Graham MJ, Crooke RM, Crooke ST, Witztum JL, Stroes ES, Tsimikas S. Antisense oligonucleotides targeting apolipoprotein(a) in people with raised lipoprotein(a): two randomised, double-blind, placebo-controlled, dose-ranging trials. *Lancet Lond Engl*. 2016 Nov 5;388(10057):2239–53.
 88. Benn M, Watts GF, Tybjaerg-Hansen A, Nordestgaard BG. Familial hypercholesterolemia in the danish general population: prevalence, coronary artery disease, and cholesterol-lowering medication. *J Clin Endocrinol Metab*. 2012 Nov;97(11):3956–64.
 89. Bell DA, Kirke AB, Barbour R, Southwell L, Pang J, Burrows S, Watts GF. Can patients be accurately assessed for familial hypercholesterolaemia in primary care? *Heart Lung Circ*. 2014 Dec;23(12):1153–7.
 90. Kwok S, Pang J, Adam S, Watts GF, Soran H. An online questionnaire survey of UK general practitioners' knowledge and management of familial hypercholesterolaemia. *BMJ Open*. 2016 Nov 1;6(11):e012691.

Köszönetnyilvánítás

Köszönöm Prof. Dr. Balla Józsefnek, a DE KK Belgyógyászati Klinika jelenlegi igazgatójának és Prof. Dr. Paragh Györgynek, a DE ÁOK Belgyógyászati Intézet Anyagcsere Betegségek Tanszék vezetőjének és a kutatások megkezdésekor az Intézet igazgatójának, hogy lehetővé tették számomra az Intézetben folyó tudományos munkában való részvételt.

Emellett köszönöm Prof. Dr. Édes Istvánnak, a Debreceni Egyetem Klinikai Központ volt elnökének, hogy segített létrehozni és elindítani azt a gépi tanulás és adatbányászati kutatást, melyből a jelen dolgozatban szereplő innovatív eredmények is kinőttek.

Köszönetemet szeretném kifejezni témavezetőmnek, Dr. Fülöp Péter egyetemi docensnek, aki támogatta és irányította a tudományos munkámat, segítségemre volt az orvosi háttér pontos megértésében, a kutatások megtervezésében és a felmerülő problémák megoldásában.

Szintén kimondhatatlanul hálás vagyok Prof. Dr. Harangi Mariann egyetemi tanárnak, aki “második konzulensként” segítette a kutatást és a munkámat és folyamatosan elérhető volt, ha segítségre volt szükségem.

Köszönöm Karányi Zsoltnak az Belgyógyászati Intézet bioinformatikai szakértőjének, hogy a kezdetektől fogja kiemelt támogatója a kutatásnak és a kutatócsoport működésének, valamint hasznos tanácsokkal és az ismereteim folyamatos bővítésével segítette a kutatómunkát.

Kiemelt köszönet illeti Dr. Daróczy Bálint mindennapos mentorálását és batáti segítségét, akinek hatalmas gépi tanulás tudása és folyamatos elérhetősége nélkül a jelen disszertáció nem készülhetett volna el.

Köszönetemet szeretném kifejezni Dr. Lőrincz Hajnalkának, valamint a DE ÁOK Belgyógyászati Intézet Anyagcsere Betegségek Tanszék Kutatólaboratórium dolgozóinak és a tudományos közleményeim társszerzőinek, akik segítették és támogatták tudományos munkámat.

Hálával tartozom családomnak és barátaimnak, akik türelmesek voltak akkor is, amikor a tudományos kutatás és a disszertáció írása során kevesebb időt tudtam velük tölteni, mint szerettem volna.

Függelék



Nyilvántartási szám: DEENK/26/2024.PL
Tárgy: PhD Publikációs Lista

Jelölt: Németh Ákos

Doktori Iskola: Egészségtudományok Doktori Iskola

A PhD értekezés alapjául szolgáló közlemények

1. **Németh, Á.**, Daróczy, B., Juhász, L., Fülöp, P., Harangi, M., Paragh, G.: Assessment of associations between serum lipoprotein (a) levels and atherosclerotic vascular diseases in Hungarian patients with familial hypercholesterolemia using data mining and machine learning.
Front. Genet. 13, 1-10, 2022.
DOI: <http://dx.doi.org/10.3389/fgene.2022.849197>
IF: 3.7
2. **Németh, Á.**, Harangi, M., Daróczy, B., Juhász, L., Paragh, G., Fülöp, P.: Identifying Patients with Familial Chylomicronemia Syndrome Using FCS Score-Based Data Mining Methods.
J Clin Med. 11, 1-14, 2022.
DOI: <http://dx.doi.org/10.3390/jcm11154311>
IF: 3.9
3. Paragh, G., Harangi, M., Karányi, Z., Daróczy, B., **Németh, Á.**, Fülöp, P.: Identifying patients with familial hypercholesterolemia using data mining methods in the Northern Great Plain region of Hungary.
Atherosclerosis. 277, 262-266, 2018.
DOI: <http://dx.doi.org/10.1016/j.atherosclerosis.2018.05.039>
IF: 4.255

További közlemények

4. Kovács, B., **Németh, Á.**, Daróczy, B., Karányi, Z., Maroda, L., Diószegi, Á., Harangi, M., Pál, D.: Assessment of Hypertensive Patients' Complex Metabolic Status Using Data Mining Methods.
J. Cardiovasc. Dev. Dis. 10 (8), 1-14, 2023.
DOI: <http://dx.doi.org/10.3390/jcdd10080345>
IF: 2.4 (2022)





5. Kovács, B., **Németh, Á.**, Daróczy, B., Karányi, Z., Maroda, L., Diószegi, Á., Nádró, B., Szabó, T., Harangi, M., Páll, D.: Determining the prevalence of childhood hypertension and its concomitant metabolic abnormalities using data mining methods in the Northeastern region of Hungary.
Front. Cardiovasc. Med. 9, 1-10, 2023.
DOI: <http://dx.doi.org/10.3389/fcvm.2022.1081986>
IF: 3.6 (2022)
6. Sturm, Á., Sharma, H., Bodnár, F., Aslam, M., Kovács, T., **Németh, Á.**, Hotzi, B., Billes, V., Sigmond, T., Tátrai, K., Egyed, B., Téglás-Huszár, B., Schlosser, G., Charnpilas, N., Ploumi, C., Perczel, A., Tavernarakis, N., Vellai, T.: N6-Methyladenine Progressively Accumulates in Mitochondrial DNA during Aging.
Int. J. Mol. Sci. 24 (19), 1-19, 2023.
DOI: <http://dx.doi.org/10.3390/ijms241914858>
IF: 5.6 (2022)
7. Alimohammadi, S., Péntes, Z., Horváth, D., Gyetvai, Á., Bácsi, A., Kis, G., **Németh, Á.**, Arany, J., Oláh, A., Lisztes, E., Tóth, I. B., Bíró, T., Szöllősi, A. G.: TRPV4 activation increases the expression of CD207 (Langerin) of monocyte-derived Langerhans cells, without affecting their maturation.
J. Invest. Dermatol. 143 (5), 801-811, 2023.
DOI: <http://dx.doi.org/10.1016/j.jid.2022.10.022>
IF: 6.5 (2022)
8. Paragh, G., **Németh, Á.**, Harangi, M., Banach, M., Fülöp, P.: Causes, clinical findings and therapeutic options in chylomicronemia syndrome, a special form of hypertriglyceridemia.
Lipids Health Dis. 21 (1), 1-14, 2022.
DOI: <http://dx.doi.org/10.1186/s12944-022-01631-z>
IF: 4.5

A közlő folyóiratok összesített impakt faktora: 34,455

A közlő folyóiratok összesített impakt faktora (az értekezés alapjául szolgáló közleményekre): 11,855

A DEENK a Jelölt által az iDEa Tudóstérbe feltöltött adatok bibliográfiai és tudományos metrikai ellenőrzését a tudományos adatbázisok és a Journal Citation Reports Impact Factor lista alapján elvégezte.

Debrecen, 2024.01.26.



Táblázatok és ábrák jegyzéke

Táblázatok

1. táblázat	A diszlipidémia Fredrickson-féle klasszifikációja és gyakoriságok	12.oldal
2.táblázat	DLCNS pontrendszer az FH meghatározásához	14.oldal
3.táblázat	A familiáris kilomikronémia szindróma pontrendszere Moulin és mtsai. alapján	17.oldal
4.táblázat	A bináris klasszifikáció igazságmátrixa	21.oldal
5.táblázat	A bináris klasszifikáció legfontosabb teljesítmény metrikái	21.oldal
6.táblázat	Főbb adathibák a vizsgált kórházi adatokban	28.oldal
7.táblázat	Klasszifikációs teljesítmény metrikák az FCS betegeken tanított modellek esetén	33.oldal
8A.táblázat	Az életkor, klinikai és laboratóriumi paraméterek szerint csoportokba sorolt betegek száma	35.oldal
8B.táblázat	Az életkor, klinikai és laboratóriumi paraméterek szerint csoportokba sorolt betegek lipidparaméterei	36.oldal
9.táblázat	A Dutch Lipid Clinical Network Score kritériumai szerint kategorizált betegek lipidparaméterei	37.oldal
10.táblázat	Az észak-alföldi egészségügyi szolgáltatókat felkereső betegek FCS pontszámai	39.oldal
11.táblázat	Az FCS pontszám becslése kulcsjellemzők alapján	41.oldal
12.táblázat	FCS pontérték kalkuláció az összes egészségügyi intézményben megjelenő betegre	42.oldal
13.táblázat	A kórelőzmény egyes tulajdonságainak jelentősége az FCS diagnosztikában, valamennyi tanított modell eredményeinek a felhasználásával	44.oldal
14.táblázat	A legmeghatározóbb laboratóriumi paraméterek határértékei az FCS diagnosztikában.	46.oldal
15.táblázat	A vizsgált populáció klinikai és laboratóriumi adatai. Az adatok mediánként (alsó-felső kvartilis) vannak megadva, n.s.: nem szignifikáns eltérés az FH és nem FH csoport között	47.oldal
16.táblázat	A kardiovaszkuláris kockázati tényezők prevalenciája FH és nem FH betegekben, valamint különböző ateroszklerotikus	49.oldal

	eredetű megbetegedésekben és mélyvénás trombózisban	
17.táblázat	A kardiovaszkuláris kockázati tényezők prevalenciája a teljes vizsgálati populációban, FH betegekben, illetve alacsony és magas (>500 mg/dl) Lp(a) szinttel rendelkező FH betegekben.	50.oldal
18.táblázat	Az egyes kockázati tényezők hatása a kardiovaszkuláris rizikókockázatra (relatív házárdok)	52.oldal

Ábrák

1A.ábra	Az exogén lipid anyagcsere ciklus folyamata	10.oldal
1B.ábra	Az endogén lipid anyagcsere ciklus folyamata	11.oldal
2.ábra	Az FH külső jegyei	14.oldal
3.ábra	Az FCS leggyakoribb klinikai tünetei	16.oldal
4.ábra	Egészségügyi intézményben megjelenők és a regionális népesség demográfiai adatai az észak-alföldi régióban	38.oldal
5.ábra	Az FCS pontrendszer gyors becslésének folyamata	40.oldal
6.ábra	A szérum Lp(a) szintek a két csoportban (FH betegek és nem FH-s egyének)	49.oldal
7.ábra	Az élet során bekövetkezett kardiovaszkuláris események aránya (%) a különböző szérum Lp(a) szinttel rendelkező csoportok szerint kategorizálva, ismert Lp(a) értékkel rendelkező FH és nem FH betegben	53.oldal