

Doktori (PhD) értekezés tézisei

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

Németh Ákos

Témavezető: Dr. Fülöp Péter



DEBRECENI EGYETEM
Egészségtudományok Doktori Iskola

Debrecen, 2024

Ritka anyagcsere betegségek diagnosztikája adatbányászat és gépi tanulás segítségével

Értekezés a doktori (PhD) fokozat megszerzése érdekében
az egészségtudományok tudományágban

Írta: Németh Ákos okleveles közgazdász, vegyészmérnök

Készült a Debreceni Egyetem Egészségtudományok doktori iskolája
(Anyagcsere és endokrin betegségek megelőzése és kontrollja programja) keretében

Témavezető: Dr. Fülöp Péter

Az értekezés bírálói:

Dr. Benczúr András, PhD
Dr. Kolozsvári Rudolf, PhD

A bírálóbizottság:

elnök: Prof. Dr. Balázs Margit, az MTA doktora
tagok: Dr. Márk László, PhD
Dr. Fiatal Szilvia, PhD
Dr. Benczúr András, PhD
Dr. Kolozsvári Rudolf, PhD

Az értekezés védésének időpontja: Debreceni Egyetem ÁOK, Belgyógyászati Intézet „A”
Épület tanterme
2024. június 5. 13:00

Bevezetés

Az egészségügyi statisztikák alapján Európában a kardiovaszkuláris megbetegedések (cardiovascular diseases, CVD) jelentik a vezető halálokat, ugyanakkor a szív-érrendszeri okokra visszavezethető korai halálozás Közép-Európában háromszor gyakoribb, mint Nyugat-Európában. A hiperlipidémia a CVD-k egyik fő kockázati tényezője, ugyanis a magas koleszterin- és trigliceridszint hozzájárul az ateroszklerózis kialakulásához már fiatal korban is. A lipidanyagcserét érintő, jelentősen emelkedett lipidszintekkel jellemezhető veleszületett betegségek, például a familiáris hiperkoleszterinémia (FH) vagy a familiáris kilomikronémia szindróma (FCS) ezért jelentős kardiovaszkuláris kockázatot hordozhatnak.

Jelen dolgozatomban a lipidanyagcsere veleszületett betegségeire, azon belül is az FH és az FCS diagnosztikai kérdéseire fókuszálok. Mindkét betegség kiemelt fontosságú abban a tekintetben is, hogy megfelelő diagnózis hiányában a lipidcsökkentő kezelés hatékonysága kifejezetten csökken ezekben, az eleve is nehezen kezelhető kórállapotokban.

Az FH az egyik leggyakoribb monogén öröklődésű betegség, melyet javarészt az low-density lipoprotein (LDL-) receptor (LDLR), vagy az apolipoprotein (Apo) B100 (Apo-B100) génjének funkcióvesztő, ill. a proprotein konvertáz szubtilizin/kexin type 9 (PCSK9) génjének funkciónyerő mutációi okoznak. Korábbi adatok alapján a betegség heterozigóta formája körülbelül 1:500, míg a homozigóta forma 1:1.000.000 gyakorisággal alakul ki. Az utóbbi évek adatai alapján a betegség valós prevalenciája ettől gyakoribb, a heterozigóta forma előfordulási gyakorisága kb. 1:300, míg a homozigóta formáé 1:300.000 körülire tehető. A világon így 20-30 millió FH beteg lehet, azonban legalább 80%-uk nem is diagnosztizált. Megemlíteném az is, hogy 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 az FH betegek átlagosan 15 évvel korábban halnak meg, mint a génhibát nem hordozó egyének.

Az éhomi kilomikronémia oka monogén öröklődésű ritka betegség is lehet. Ennek során a csökkent lipoprotein lipáz (LPL) aktivitás miatt károsodik a trigliceridekben gazdag lipoproteinek keringésből történő eliminációja. Ez az állapot súlyosan magas – és tartós – szérumszintű triglicerid (TG) koncentrációhoz vezet FCS-ben, mely súlyos, potenciálisan életveszélyes állapothoz vezethet recidív hasnyálmirigy-gyulladással. Becslések szerint az FCS egymillió emberből egyet érint és minden esetben az LPL fehérje működését jelentős mértékben károsító genetikai eltéréssel jár akár homozigóta, akár összetett (compound) heterozigóta mutációk formájában. Bár az FCS betegek döntő többségében funkcióvesztő

homozigóta mutáció mutatható ki az LPL génjében, ugyanakkor 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 gliceraldehyd-3-foszfát dehidrogenáz (G3PDH1) génjeiben is.

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 és időigényes, valamint speciális szakorvosi tudást igényel. Fentiek alapján célul tűztük ki, hogy az ún számítógéppel támogatott diagnosztika és terápia segítségével régiókban megbecsüljük ezen betegségek pontos előfordulását és olyan hatékony diagnosztikai eszközt adjunk az orvosok kezébe, amely különleges szakértelem nélkül is segítséget adhat a betegségek pontos felismerésében, ill. ezen kórállapotok hátterét és dinamikáját jelző okok és kockázati tényezők feltárásában, valamint a terápiát befolyásoló faktorok összegyűjtésében.

Exogén és endogén lipidanyagcsere

A koleszterin 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 trigliceridek a szervezet fő energiaforrásként 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 lipidanyagcsere két fő folyamatra, az ún. exogén és endogén lipidanyagcserére bontható. Az exogén ciklus folyamata során a táplálkozással bevitt zsír emésztést követően kilomikronokká alakul a vékonybél hámsajtjeiben. 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 részecskéket 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, 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.

Diszlipidémiák

Etiológiai szempontból, a diszlipidémiák primer (valamilyen genetikai eltérés által meghatározott, gyakran familiáris halmozódást mutató) és szekunder (azaz egyéb betegséghez, vagy gyógyszereszedéshez 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.

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. 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. Az utóbbi időben egyre nagyobb jelentőséget kapnak az adatbányászattal 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 szérumban lipidparaméterek, így a koleszterin, TG, HDL-C és LDL-C keringő szintjeinek mérése.

A Fredrickson IIA típusú diszlipidémiát eredményező FH igen magas koleszterin szinttel, azon belül is jelentősen emelkedett LDL-C értékkel jellemezhető. A szövetekben lerakódó koleszterin miatt a betegek 20-40%-ánál megjelenik valamilyen patognomikus elváltozás is, mint pl. az Achilles-inak tendon xanthomája, a szemhéjak xanthelasmája, vagy a szaruhártya gyűrűs elszíneződése (arcus corneae). Az FH diagnózisában hatékony segítséget nyújt az ún. Dutch Lipid Clinic Network (DLCN) kritérium és pontrendszer, 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.

A Fredrickson 1-es típusú diszlipidémiával jellemezhető FCS-ben 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 szérumban CM szint máj- és lépmegegyesítéshez vezethet. Az FCS legsúlyosabb potenciális szövődménye ugyanakkor a heveny, nem ritkán ismétlődő hasnyálmirigy gyulladás, melynek mortalitása elérheti az 5-6%-ot. A diagnosztikát támogató pontrendszer FCS esetén is rendelkezésre áll, melyet Moulin és szerzőtársai javasoltak (FCS score).

A fent javasolt rutin lipidparaméterek mellett az újabb ajánlások az apoB100

gyakoribb mérését is javasolják, mivel ennek szérumszintje arányos a VLDL, LDL és az IDL értékeivel és a kardiovaszkuláris kockázat jó jelzője. Szintén indokolt a szérumban lipoprotein(a) [Lp(a)] legalább egyszeri meghatározása, mely az ASCVD független rizikófaktora, ugyanakkor a szokványos lipidcsökkentő kezeléssel nehezen befolyásolható. Az Lp(a) és az FH betegek morbiditása, ill. mortalitása közötti összefüggések pedig a mai napig nem teljesen tisztázottak az irodalmi adatok alapján.

Szöveges adatok gépi feldolgozása

A szövegbányászat (text mining), mely az adatbányászat egy külön ága, mely a szöveges adatok feldolgozásával foglalkozik és magában foglalja a számítástechnika és -tudomány, 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) során a nyers, rendezetlen szöveghez strukturált információt rendelünk, a folyamat fontos lépései a "tokenizáció", a szűrés (filtering) és a "stemming" (szótővezés), melyek a rendezetlen szövegek előfeldolgozásának járulnak hozzá. A szövegbányászat fontos eleme még az ún. "regexek", azaz szabályos kifejezések (regular expressions) alkalmazása, melyek lehetőséget adnak szabályok és mintázatok egyszerű leírására, így alkalmazhatóak a már előfeldolgozott strukturálatlan szövegeken, egyszerűbb és összetettebb kifejezések keresésére is.

A dokumentumok és kifejezések rangsorolása egyike a legfontosabb szövegbányászati feladatoknak. Ennek egyik legtöbbet használt technikája, a TF-IDF (Term Frequency - Inverse Document Frequency), amely a dokumentumokban használt szavak, kifejezések előfordulási gyakoriságát vizsgálva következtet azok fontosságára, relevanciájára. 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).

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 struktúrájának a reprodukálását célozzák és ezáltal alkalmasak szövegek összetett morfológiai, szintaktikai (mondattani) és szemantikai (jelentéstani) elemzésére. Az utóbbi időkben leginkább a neurális hálózatokat használják, de léteznek statisztikai és szabályalapú megközelítések is.

A becsült és a tényleges előfordulási valószínűség közötti eltérést perplexitásnak nevezzük, ami a nyelvi modellek egyfajta teljesítménymetrikájának is felfogható. Azaz, minél alacsonyabb a perplexitás, annál sikeresebb egy nyelvi modell az adott szöveg struktúrájának megtanulásában. 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 word2vec modelleket a transzformer alapú modellek haladták túl a 2010-es közepén, melyek a neurális hálókra épülő nyelvi modellek közül jelenleg a legsikeresebbek. A transzformer alapú modellek (pl. BERT, GPT) 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 elosztott cloud rendszereken.

Bináris klasszifikáció

A bináris klasszifikáció során a tárgyat, azaz a bemeneti mintát 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). A klasszifikáció matematikailag-statisztikailag tulajdonképpen a regressziószámítás speciális esetének is tekinthető (ahol a “függő” változó mindig diszkrét értékeket vehet csak fel). Az osztályozás során az input adatokból, tanuló algoritmusok segítségével egy klasszifikációs modellt építünk, amely a lehető legpontosabban képes leírni az $(\mathbf{x})$ és (y) közötti kapcsolatot (tréning halmaz), valamint szintén képes helyesen kategorizálni olyan adatokat, amelyek nem voltak részei a modellépítésnek (teszt halmaz). A bináris klasszifikáció talán legfontosabb és legismertebb összetett teljesítménymetrikája a ROC-AUC görbe. 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 klasszifikációs modellek (“classifier”-ek), azaz az osztályozó algoritmusok széles skálája érhető el. A **döntési fák** (klasszifikációs vagy regressziós fák) az egyszerűbb osztályozó modellek közé tartoznak. 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). Az **NN klasszifikáció** (“k-nearest neighbors” classifier) a legjelentősebb “lazy learner”. A “lazy learner” (szemben az eddigi “eager learner” modellekkel) azt jelenti, hogy a betanítási adatok általánosítása nem történik meg a klasszifikáció előtt. Ez különösen a nagyméretű és gyakran változó adathalmazok esetén hasznos. Az NN klasszifikáció során azt vizsgáljuk, hogy egy d-dimenziós állapottérben egy vizsgált x pont környezetében található k darab eset milyen kategóriákba tartozik. 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. Ez legtöbbször egy nagyon erős megkötés, hiszen a klasszifikációk nagy részénél kisebb-nagyobb korreláció megfigyelhető. A **mesterséges neurális hálózatok** (ANN, artificial neural networks) pedig a gépi tanulás legnépszerűbb és leggyorsabban fejlődő modellei. Egy ANN csomópontokból (“neuronokból”) és a csomópontokat összekötő kapcsolatokból áll (hasonlóan az idegrendszer neuron-axon-dendrit kapcsolataihoz). A neuronok aktiválása aktiváló függvények segítségével történik, melyek meghatározzák az ANN típusát. 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 kapcsolnak. 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, valamint jelentős tanítási idővel jár. Az **SVM** (support vector machine, vagy support vector network), mely sekély neurális hálónak is tekinthető, a sekély tanulási módszerek talán legnépszerűbbike. 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. Végül 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.

Célkitűzések

A ritka anyagcserebetegségek 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

Vizsgálati populáció

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, anonimизált adatot használtunk. A Debreceni Egyetem Klinikai Központban és a Nyíregyházi Jóna 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, anonimизált adatbázist használtunk.

Felhasznált adatok

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. Alapvetően 5 típusú klinikai adattal kerültünk kapcsolatba a kutatásaink során, melyek az alábbiak voltak: a betegvizsgálat szöveges adatai (anamnézis, epikrízis, műtéti leírás), a laboratóriumi vizsgálatok numerikus adatai és egyéb fontos paraméterei (pl. a referenciatartományok, mértékegységek, a vizsgálatokat végző gépek adatai), a feltételezett és igazolt diagnózisok adatai, a beavatkozások és vizsgálatok eredményei, valamint 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, melyek az adat természetesen velemjárói. 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 meglehetősen nehezíti a tisztánlátást és sokszor részletesebb elemzéseket és mélyebb feltárásokat igényel.

A tisztítási és előfeldolgozási folyamatok részben nyílt forrású eszközök segítségével (Pandas, Numpy), részben saját fejlesztésű Python 2.8.X és Python 3.X szkriptek segítségével történtek. Végeredményként nagy, összefüggő és indexelt adatfájlokat kaptunk, melyek minden orvosi eset diagnózisát, laboratóriumi vizsgálatát, a beavatkozások és diagnosztikai eljárások leírásait, valamint az összes szöveges adatot tartalmazták. Az egyének személyi összefüggő anonim statisztikai adatai külön táblában kerültek tárolásra. Az előfeldolgozási

folyamat során különféle szerializálási és pufferekési módszereket is alkalmazni kellett a masszív adattömeggel kapcsolatos memória problémák kiküszöbölésére. Ezen feladatokat elsősorban a python alapú pandas, a unix környezetben népszerű awk nyelv, ill. dask framework-ök segítségével oldottuk meg. Az előfeldolgozási folyamatok végén az adat PostgreSQL 13.x adatbázis táblákba került rögzítésre.

Adattisztítás adatbányászattal

A szöveges adatok strukturálásához és a bennük szereplő információ kinyeréséhez további feldolgozási lépések szükségesek. A fontos kifejezések megtalálásához és csoportosításához parsing, stemming, stop word szűrés, BOW (bag-of-words) modellezés és kifejezések rangsorolása (TF-IDF, W2V) történt. 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 többmillió kifejezésteret kétszázötvenezerre 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) 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, valamint az általános kifejezések esetében a BERT (Bidirectional Encoder Representations from Transformers) modellt használtuk. 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.

Páciensek szűrése FH és FCS diagnózisra

A DLCN score (DLCNS) meghatározása előtt kiszámoltuk a betegek átlagos lipidértékeit, 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 tisztított és előfeldolgozott adatokon hasonlóan történt az FCS score- 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.

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. Ahhoz, hogy az egyes gépi tanulás 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 teljesítménymetrikák alapján (ROC-AUC, ACC, SENS) 50-50%-ban tanuló- és teszhalmazra bontva az adatot. A Vapnik-Chervonenkis féle generalizációs elmélettel összhangban három – széles körben alkalmazható – matematikai modellt használtunk, melyek az ensemble fák (tree ensembles), a “sekély” kernel alapú (SVM) és “mély” ReLU aktivációs neurális hálók. Mivel az FCS diagnózisok esetén a struktúrára semmilyen speciális feltétel nem teljesül, így teljesen kapcsolt (“fully connected”, “struktúra-agnosztikus”) neurális hálókkal dolgoztunk, ReLU aktivációs függvényrel és adaptív momentum (Adam) optimalizációval.

Az FCS diagnózisok kis adathalmazok, melyeknél az adatmátrix tulajdonságok sorrendisége nem meghatározó, így már elsőre is valószínűsíthető volt, hogy az ensemble fák és a kernel-alapú modellek lesznek hatékonyabbak az osztályozás során. A legfontosabb teljesítménymetrikák bináris osztályozásához a szenzitivitást (azaz, a tényleges FCS betegek hány százalékát mondja a modell FCS betegnek), a specifitást (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. 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. Az AdaBoost, SVM implementáció során scikit-learn 0.23, míg a ReLU hálókhoz PyTorch 1.6, a fákhoz pedig XGBoost 1.2.1 programokat alkalmaztunk.

Laboratóriumi vizsgálatok

A vizsgálatok mindig az adott egészségügyi intézmény aktuális protokolljának megfelelően történtek, nem feltétlenül éhgyomorra. Az Lp(a) szintek tizenkét órás éhezést követően, éhgyomorra levett vérmintából 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.

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. 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).

Eredmények

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 szolgáltathat 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. Az első csoportba kerültek azok, akiknél biztos FH diagnózis volt felállítható a szöveges adatbányászat segítségével, míg a másodikba soroltuk azokat, 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.

Eredményeink alapján a régió betegeinek 0,89%-át érintheti a familiáris hiperkoleszterinémia. Mivel Magyarországon a teljes lakosság közel 76%-a átlagosan évente egyszer megfordul az állami egészségügyben és a vizsgált populáció epidemiológiai adatai lényegében egyeznek a 2011-es regionális népszámlálási adatokkal, a klinikai prevalenciát feltehetően mérvadónak tekinthetjük a tényleges prevalencia becslése során. A három vagy a

feletti DLSN értékkel rendelkező pácienseknek felajánlott genetikai vizsgálat az irodalomban 33%-ban erősítette meg genetikailag az FH diagnózist. Mindezek alapján, számításaink szerint hazánkban minden 340 emberből 1 szenvedhet familiáris hiperkoleszterinémiában (11.937 beteg rendelkezett 3 feletti DLSN értékkel, ezek 33%-a 3940 beteg, ami körülbelül 340-ed része a teljes betegpopulációnak).

Az értékelésünk nem vette figyelembe a DLNCS genetikai és családi anamnézisre vonatkozó elemeit, mivel ezek az adatok csak rendkívül korlátozottan voltak elérhetőek az adatbázisban. 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. A legjobb tanulás esetén feltételeztük, hogy a csoport a lehető legjobban különül el a többiektől. 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 FH-ban szenved.

A familiáris kilomikronémia szindróma esetén is kiszámoltuk minden beteg FCS pontszámát. 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 (19,4 ppm).

Az FCS prevalencia gyors becslését máshogy is megközelítve olyan, ún. kulcsjellemzőket próbáltunk meghatározni, melyek a nagy esetszámot gyorsan nagyságrendekkel alacsonyabb esetszámmra csökkentik ("vágják"), hasonlóan az FH vizsgálata során is alkalmazott módszerhez. Ritka anyagcserebetegségeknél gyakori egy-egy laboratóriumi érték 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 gyors becsléshez magas FCS pontszámú betegeket választottunk ki. A betegség kulcsjellemzőinek kiválasztásakor nagy pontszámmal beleszámító, valamint jól mérhető elemeket használtunk. Ilyen kulcsjellemző volt a (i) 10 mmol/l-t meghaladó szérumszint

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. Az FCS pontszámot az adatbázisban hozzáférhető minden egyes betegre kiszámolva 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.

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 a valószínűtlen és nagyon valószínű FCS besorolást kapó, a Debreceni Egyetem Klinikai Központjában megjelent 275 beteg adatát részletes szakorvosi ellenőrzésnek vetjük alá. Így 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, hasonlóan az adatbányászati eredményekhez.

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. Másrészt, a feltárt összefüggések a betegség mélyebb orvosi összefüggéseinek feltárásában is jelentős segítséget nyújthatnak. A kutatásaink során mindkét irányban végeztünk elemzéseket, így az FH betegek esetén a gépi tanulást arra használtuk, hogy minél homogénebb betegségcsoportokat képezve pontosan azonosíthassuk a DEKK-ben megjelent FH betegeket, azaz a tényleges betegek tulajdonságain keresztül minél jobban megértsük az FH-ban szenvedő egyének egyéb tulajdonságait.

Az debreceni FCS betegek adatain, mint vektortereken tanítva, többféle gépi tanulási módszerrel matematikai modelleket építve további lehetséges betegeket kerestünk mindkét egészségügyi intézményben. Tréninghalmazként a fenti 21 beteg adatát használtuk, míg teszthalmazként a 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 tegyünk.

Megfigyeléseink megerősítették, hogy az egyes laboratóriumi paraméterek relatív fontosságának vizsgálatakor a TG szintek a legfontosabb diagnosztikai kritériumok, hiszen az FCS is leginkább a szérum triglicerid értékekkel karakterizálható. 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 kórelőzményben szereplő átlagos és legmagasabb koleszterin szintek szintén alapvetőnek bizonyultak abból a szempontból, hogy az FCS betegek a nem FCS páciensektől megkülönböztethetőek legyenek. Ezen megfigyelésünk különösen érdekes annak tükrében, hogy az FCS diagnózisának felállítása során a koleszterin értékek vizsgálata jelenleg nem része a klinikai gyakorlatnak.

További érdekes összefüggéseket találtunk, amikor megvizsgáltuk, hogy mik azok a határértékek, ahol a matematikai modellek döntésében változás következik be, azaz az FCS-ből a nem FCS, illetve a nem FCS-ből az FCS diagnózisa felé mozdulnak. 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 már említett triglicerid és koleszterin határértékek mellett a hemoglobin, MCHC, bazofil granulocita, limfocita vagy amiláz eredmények megállapított 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.

FH betegek komplex vizsgálata Lp(a) kategóriák szerint

Ezen 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 betegek tekintettünk. Mindezek alapján 459 beteg volt azonosítható, akik 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) mg/l vs. 167 (80,2; 431.5) mg/l, $p < 0,01$]. Mindkét csoportban azt találtuk, hogy a nőknek szignifikánsan magasabb Lp(a) értékeik voltak, mint a férfiaknak [266 (108–875) mg/l vs. 182 (73,1–648) mg/l, $p < 0,05$] az FH betegek között és [179 (81,5–458) mg/l vs. 152 (80,7–386) mg/l, $p < 0,01$] a nem FH egyének körében. Bár az ateroszklerotikus szövődmények 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 szövődmények gyakorisága között nem találtunk szignifikáns összefüggést az FH csoporton belül.

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. 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. A kockázati tényezőket magas és alacsony Lp(a) értékkel rendelkező FH betegeknél is megvizsgálva lényeges eltérést nem találtunk a két csoport között. A nemek közötti összehasonlításban ugyanakkor megfigyelhető volt, hogy magas Lp(a) értékek minden csoportban gyakrabban fordultak elő nőknél, 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.

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. “Egészségesnek” itt az alábbi kondícióktól mentes egyéneket tekintettük: akiknél túlsúly, dohányzás, sztatinkezelés, hipertónia, diabétesz, magas Lp(a) nem volt megállapítható, emellett a normál tartományba eső LDL-C és HDL-C értékekkel rendelkeztek. 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. Az összes kockázati tényezővel halmozottan rendelkező betegek körében a relatív hazard 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.

Az új eredmények összefoglalása

- 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.
- A korábban javasolt pontrendszer adatbányászati módszerekkel történő alkalmazásával az FCS regionális gyakorisága 19,4:1.000.000-ra tehető; 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.
- 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).
- Az általunk fejlesztett gépi tanulás modellek segítségével felvethető egyes új laboratóriumi paraméterek figyelembe vétele az FCS pontrendszer további finomítása érdekében. 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 a legnagyobb az FCS diagnózisának valószínűsége (jelen vizsgálataink alapján ez a sáv 6,5 - 11,0 mmol/l között van).
- Az FH betegek részletes vizsgálata során az ASCVD 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 hazard értékek azonosak.

- A nemi különbségeket tekintve, a szérumban Lp(a) szintek a nők között FH-ban és FH-val nem érintett populációban is szignifikánsan magasabbnak bizonyultak.
- Az 1000 mg/l feletti Lp(a) értéknél a kardiovaszkuláris rizikó szignifikánsan megnő. Az efeletti érték relatív hazardja 1,6-6,4 közötti, mely bár jelentős, de mégis alacsonyabb a gyakori másodlagos tényezők (így a dohányzás, vagy a hipertónia) egyéni kockázatához képest.
- Összességében, 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. Ezen betegségek pontos gyakoriságának meghatározása is igen nagy kihívást jelent. A genetikai predispozíciótól függően, a familiáris hiperkoleszterinémia 1:250-1:1.000.000 gyakorisággal fordul elő a populációban, míg a familiáris kilomikronémia szindróma prevalenciája szintén 1:1.000.000-ra tehető. 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. Az FH szűrésének kiemelt szerepe van a kardiovaszkuláris megbetegedések megelőzésében, emellett az újonnan fejlesztett és hatékony lipidcsökkentő terápiák (pl. a PCSK9-gátlók) alkalmazása szintén hozzájárulhat az FH szűrések felgyorsulásához, mivel egyre több beteg helyzete javítható ezekkel az új terápiás lehetőségekkel. A debreceni betegek adatain később elvégzett gépi tanulási módszerekkel tovább finomítva az FH betegek azonosítását 459 valószínűsíthetően FH beteget találtunk, 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 alulbecsülhattuk a tényleges prevalenciát.

Az FCS-t hasonló módszerekkel vizsgálva a betegség gyakoriságát 19,4 ppm-nek találtuk, ami nagyobb, mint a becsült 1 ppm globális prevalencia. 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. az FCS prevalenciája 13 ppm-nek adódott egy hasonló nagyságú, szintén egészségügyi szolgáltatók által nyújtott ellátást igénybevevő populáción. Gyermekekben, kisebb betegtömeget nézve ez a szám 1:300.000-nek adódott, ami szintén magasabb prevalenciát sejtet, mint a jelenleg általánosnak elfogadott becslések.

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 a leginkább specializált 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 betegtömegben alkalmazott módszertan (adatbányászat és gépi tanulás) is alátámaszt.

Az FCS diagnosztikája jelenleg nagyban támaszkodik a csak a központokban elérhető genetikai analízisre, valamint az LPL enzimaktivitás mérésére. A közelmúltban ugyanakkor fontos változást hozott a lipidológus szakértők által alkotott FCS score, mely az FCS gyors és hatékony felismerését támogatja. Az FCS pontrendszer alapvetése a szérum triglicerid szint ismerete, mely minden ellátási szinten elérhető laboratóriumi paraméter. Ahogy mi is igazoltuk, csupán a szérum 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 vizsgálandó populáció gyorsan 1-2%-re szűkíthető. Ez a szoros és nyilvánvaló kapcsolat a szérum 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, amely szükséges a pontszám biztos kiszámításához. Mindezek 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 szolgálhat, különösen ha nagy adathalmazokban, automatizált módszerekkel próbáljuk a potenciális FCS betegeket azonosítani.

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 betegek szétválasztására, ugyanakkor a neurális hálók lényegesen rosszabbul teljesítettek. Emellett bizonyos laboratóriumi paraméterek elvételével és hozzáadásával újabb tanításokat végeztünk, másrészt a már sikeresnek bizonyult modellekbe változó értékeket táplálva próbáltuk megtalálni a legfontosabb szérumszinteket (ún. “vágásokat”), ahol a “beteg - nem beteg” bináris szeparáció a legjobb hatékonysággal végbemegy. Azt találtuk, hogy a szérum TG értékek vizsgálata nyilvánvalóan segíti – ill. az összes laboratóriumi érték közül a legjobban segíti – az FCS pontos diagnózist.

Véleményünk szerint érdekes és fontos újdonság, hogy a szérum összkoleszterin értékeinek vizsgálata további segítséget tud nyújtani a lehetséges FCS betegek identifikálásában, 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.

A lipidparaméterek mellett számos egyéb metabolikus komponenst is találtunk (pl. a transzamináz aktivitások, vagy a szérum 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, melyösszhangban van egy, néhány éve megjelent tanulmány megállapításaival is.

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, ill. ezeket jól definiált csoportok között összehasonlítsuk. Ilyen elemzésre, így pl. a metabolikus paraméterek eltérő szintjeinek hatásvizsgálatára ad lehetőséget a szérum Lp(a) értéke 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. Ismert, hogy a különböző kockázati tényezők együttes előfordulása jelentősen növeli az ateroszklerózis kockázatát, hozzájárulva a fenotípusos heterogenitáshoz. Bár már azonosították az Lp(a) szintje, valamint a CVD kockázat közötti kapcsolatot heterozigóta FH páciensekben, az emelkedett Lp(a) koncentrációk rizikófaktoroként való potenciális szerepe mind a mai napig nem tisztázott ezen betegségben.

Hazánkban elsőként próbáltuk meg 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. Ugyanakkor, nyugat-európai és dél-amerikai populációkon végzett felmérések során a magasvérnyomás-betegség prevalenciája ennél lényegesen kisebbnek, 11-24% közöttinek adódott FH betegekben.

Megvizsgálva a dohányosok arányát, azt találtuk, hogy hazánkban meglepően sok FH beteg dohányzik (66,4%), melyhez igen hasonló adatot találunk a lengyel betegek körében, míg más országokban ez az arány lényegesen kisebb volt. Nagyon hasonlóan lengyel, román vagy mexikói FH betegekhez, a diabétesz gyakorisága lényegesen nagyobbak bizonyult az FH betegeink körében, mint a non-FH populációban. Ezen, változatos adatok is aláhúzzák a rizikófaktorok vizsgálatának fontosságát, valamint rávilágítanak a széleskörű, nemzetközi felmérések szükségességére ebben a betegcsoportban. Érdeemes megjegyezni azt is, hogy vannak tanulmányok, amelyek sokkal kisebb diabétesz prevalenciát tapasztaltak és olyan megfigyelést is tettek, mely szerint a mutációk következményének súlyossága FH betegek esetén fordítottan arányos a diabétesz prevalenciával. 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 alapján – az FCS-hez hasonlóan – a betegek és a hozzátartozók figyelmének felhívása erre, az önmagában is komoly egészségügyi kockázattal járó állapotra, kiemelt jelentőséggel kell, hogy bírjon.

Az FH páciensek Lp(a) értékeit viszonylag kevés tanulmányban vizsgálták részletesen, ugyanakkor az eddig ismert adatok a mi eredményeinkkel jelentős összhangban vannak. Megemlítenéd az is, hogy az irodalmi adatokkal szintén egyezően, a nők körében szignifikánsan magasabb Lp(a) értékeket találtunk. Ez a különbség egyébként nem csak az

FH betegek körében figyelhető meg, de az általános populációban is. Bár a nemek közötti különbség pontos oka nem ismert, de feltételezhető, hogy az ösztrogén apo(a) fehérje expressziójára kifejtett hatása felel az eltérésért.

Bár az Lp(a) mérések alacsony száma gátolta a szignifikáns eltérések azonosítását az FH betegekben, 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 hazard kiszámításához azonban további elemzések szükségesek.

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. 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. 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. A jelenleg is folyó klinikai kutatások függvényében ezek az új hatóanyagok jelentős segítséget nyújthatnak a magas Lp(a) szérum értékekkel 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. Kétségtől 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.

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.

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, 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, 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.

További eltérést okozhat, hogy bizonyos vizsgálatok, így pl. 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.

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.

A fentiek mellett arra jutottunk, hogy a gépi tanulás eszközrendszerével maga az FCS pontrendszer is tovább fejleszthető. Mindez azért is tűnik hasznosnak, mivel a pontszám kiszámítása során olyan információkra is támaszkodni kell, amelyek 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 is, ezáltal a korai kezelés megkezdésére, a potenciálisan súlyos szövődmények elkerülésére.

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.

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 korábbi 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, a 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 baráti segítségét, akinek a gépi tanulást érintő hatalmas 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.



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., Champilas, 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 tudománymetriai ellenőrzését a tudományos adatbázisok és a Journal Citation Reports Impact Factor lista alapján elvégezte.

Debrecen, 2024.01.26.

