Okozati következtetések levonása megfigyeléses adatokból

Ferenci Tamás
tamas.ferenci@medstat.hu
http://www.medstat.hu/
https://www.youtube.com/c/FerenciTamas

Utoljára frissítve: 2022. június 30.

Kauzalitás vizsgálata megfigyeléses adatokból

- Annyit mondtunk, hogy „tenni tehetünk a confounding ellen" meg, hogy „statisztikai úton szűrhető" (hiába megfigyelésesek az adataink); de mit jelent ez?
- 5 fő módszer (számos variánssal):
  - Rétegzés
  - Standardizáció
  - Illesztés (matching)
  - Többváltozós – regressziós – modellezés
  - Propensity score eljárások
- De *bármelyiket* is használjuk, azt nem lehet megkerülni, hogy csak azokat tudjuk szűrni, amikről eszünkbe jutott, hogy confounderek, és le is tudtuk mérni őket! (Ez volt a kísérletek nagy előnye!)
- Plusz, a módszertől függően különböző további feltevések jöhetnek be, tehát ezek sem univerzálisak
- Fontos, hogy ezek a confounding analitikai fázisban történő korrigálásnak eszközei – tehetünk tervezési fázisban is lépéseket
- A manapság legfontosabb eljárás a többváltozós modellezés, azzal külön is fogunk foglalkozni; a cél most inkább az, hogy kiderüljön: egyáltalán megoldható ez a probléma

A rétegzés alapötlete

- Tulajdonképpen már láttuk:

	Nem szed HRT-t	Szed HRT-t
Összességében	2,6% (290/11000)	2,3% (72/3200)

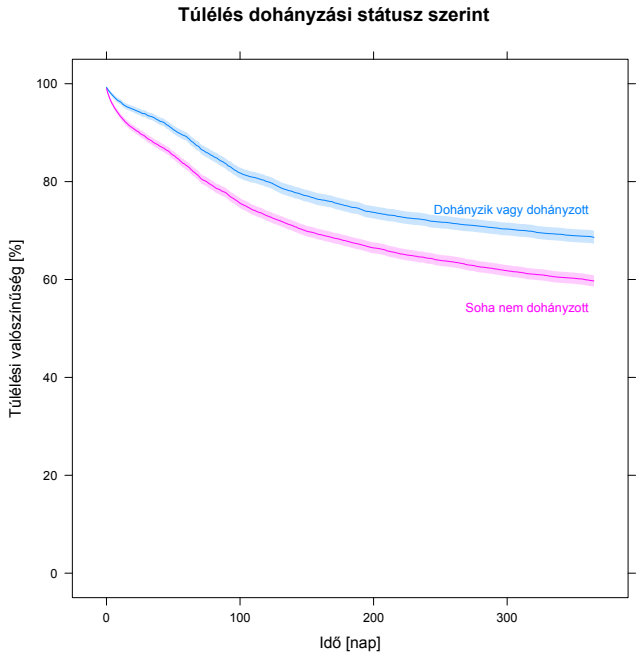
- Rétegzésnek hívják a statisztikusok, amikor az adatokat valamilyen szempont szerint hézag- és átfedésmentes csoportokra bontjuk, és mindegyikben külön-külön végrehajtjuk az elemzést:

	Nem szed HRT-t	Szed HRT-t
Alacsony szocioökonómiai státusz	4% (240/6000)	6% (12/200)
Magas szocioökonómiai státusz	1% (50/5000)	2% (60/3000)
Összességében	2,6% (290/11000)	2,3% (72/3200)

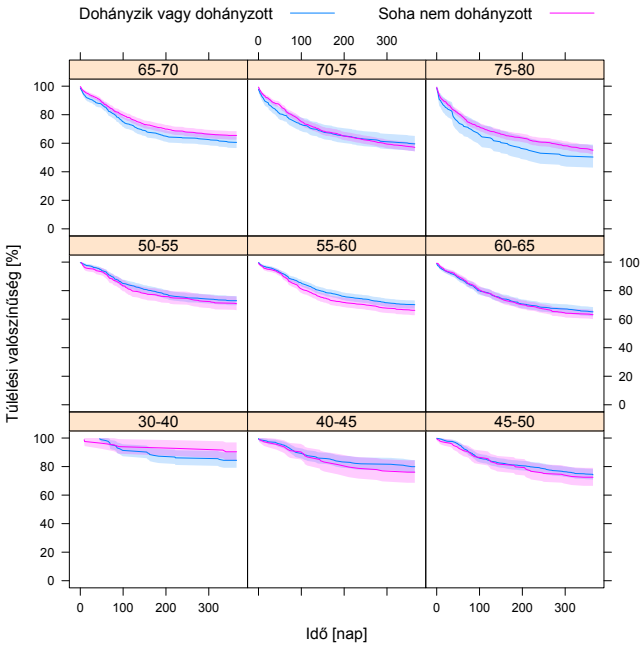
A rétegzés előnyei és hátrányai

- Előny: semmilyen feltételezéssel nem kell élni, univerzális választ ad
- Hátrány: nem egy választ ad (hanem annyit, ahány réteg van)
- Ez részint nehezen áttekinthető, részint bizonytalan (minél több réteg van, annál kevesebb alany jut egybe)
- A gyakorlatban simán lehet 10-20, vagy akár annál is több confounderünk (ne felejtsük el, hogy ezek igazából potenciális confounderek, a valóságban mi sem tudhatjuk, hogy mik az igaziak, ezért az összes potenciálisat be kell raknunk)
- Ezek egyáltalán nem biztos, hogy kétkimenetűek, de a legnagyobb gond, hogy ha teljesen általánosak akarunk lenni, akkor ezek kombinatorikusak
- Külön problémát jelentenek a folytonos változók: szét kell vágni, de azt meg nem lehet jól csinálni (ha túl széles a sáv, akkor különböző dolgokat mosunk egybe, ha túl szűk, akkor kevés alany lesz egy rétegben)

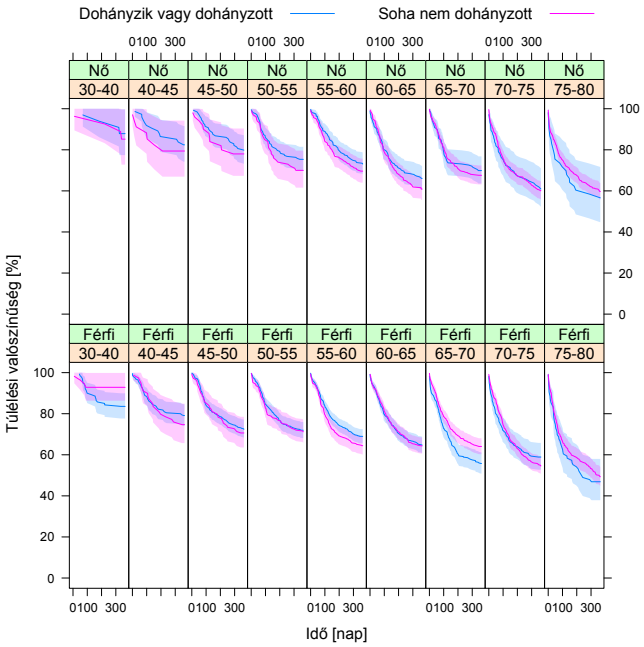
A rétegzés illusztrálása



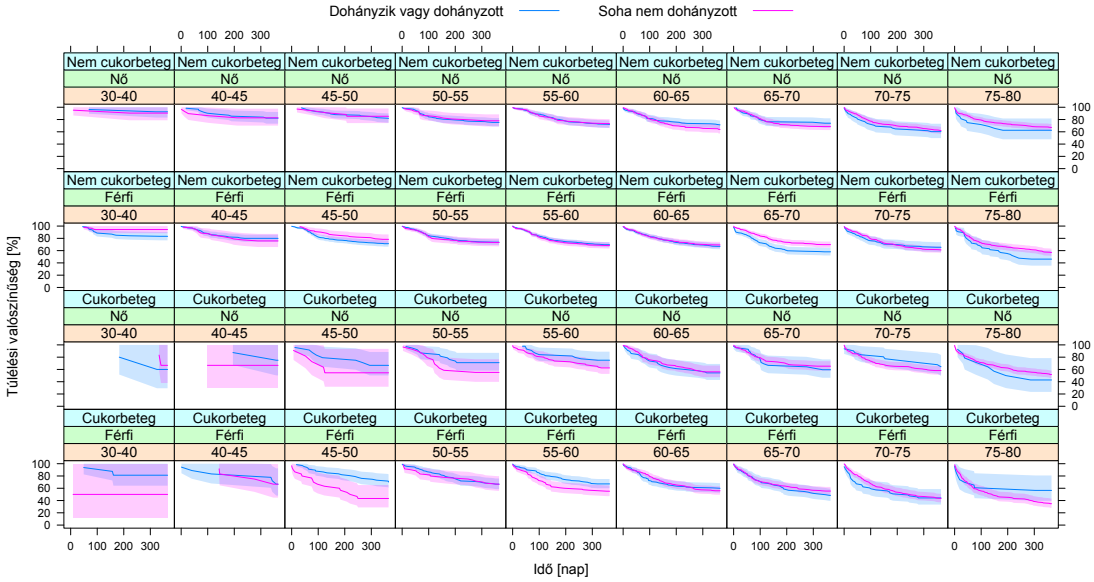
A rétegzés illusztrálása



A rétegzés illusztrálása



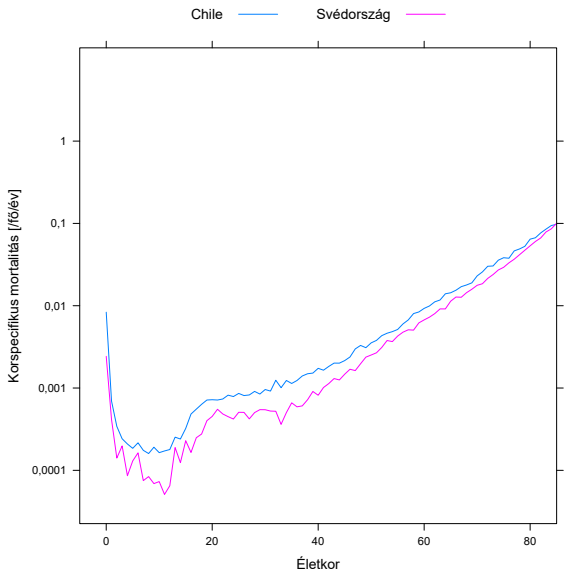
A rétegzés illusztrálása



A standardizálás illusztrálása egy problémán

- Vegyük elő a svéd-chilei példát!
- Emlékeztetőül:
- Svédországban 2005-ben 91 ezer 710 halálozás történt, a lakosságszám 9 millió 10 ezer 729, így a nyers halálozási ráta (CDR) 10,2/ezer fő/év
- Chilében ugyanabban az évben 86 ezer 100 halálozás történt, a lakosságszám 15 millió 519 ezer 347, így a nyers halálozási ráta 5,5/ezer fő/éve
- Svédországban kétszer (???) nagyobb a halandóság?

A rétegzés illusztrálása



(Igen, igazából ez egy rétegzés volt! Több eredmény van, sőt, nagyon sok, csak ügyes ábrázolással ezt mégis jól áttekinthetővé tettük!)

A számítás menete (Svédország példáján)

Korcsoport	Halálozások száma	Létszám	Létszám megoszlás	Korspecifikus mortalitás
0	378	965477	0,1071	0,0004
10	215	1192801	0,1324	0,0002
20	518	1068031	0,1185	0,0005
⋮	⋮	⋮	⋮	⋮
100	637	1228	0,0001	0,5188
	91710	9010729	1	

$$\mathrm{CDR}_{\text{Svédország}} = \frac{91710}{9010729} = \frac{378 + 215 + 518 + \ldots + 637}{9010729} =$$
$$= \frac{965477 \cdot 0{,}0004 + 1192801 \cdot 0{,}0002 + 1068031 \cdot 0{,}0005 + \ldots + 1228 \cdot 0{,}5180}{9010729} = \ldots$$

A számítás tehát (Svédország példáján)

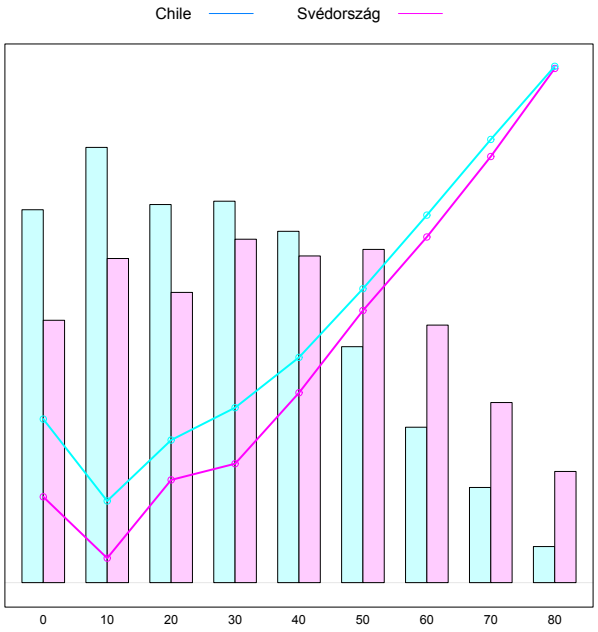
$$\mathrm{CDR}_{\text{Svédország}} = \ldots = \frac{965477 \cdot 0{,}0004 + 1192801 \cdot 0{,}0002 + 1068031 \cdot 0{,}0005 + \ldots + 1228 \cdot 0{,}5180}{9010729} =$$
$$= \frac{965477 \cdot 0{,}0004}{9010729} + \frac{1192801 \cdot 0{,}0002}{9010729} + \frac{1068031 \cdot 0{,}0005}{9010729} + \ldots + \frac{1228 \cdot 0{,}5180}{9010729} =$$
$$= \frac{965477}{9010729} \cdot 0{,}0004 + \frac{1192801}{9010729} \cdot 0{,}0002 + \frac{1068031}{9010729} \cdot 0{,}0005 + \ldots + \frac{1228}{9010729} \cdot 0{,}5180 =$$
$$= 0{,}1071 \cdot 0{,}0004 + 0{,}1324 \cdot 0{,}0002 + 0{,}1185 \cdot 0{,}0005 + \ldots + 0{,}0001 \cdot 0{,}5180 = 0{,}010178$$

Azaz azt is mondhattuk volna, hogy a korspecifikus mortalitásokat szorozzuk a korosztály létszámarányával, és ezeket összeadjuk – tehát lényegében egy súlyozott átlag: a korspecifikus mortalitások súlyozva a korcsoportok létszám szerinti megoszlásával, tehát a korfával

Emlékeztetőül a magyarázat

- A kutya ott van elásva, hogy nagyon mások a „súlyozó függvények", tehát a korfák!
- Igaz ugyan, hogy minden életkorban rosszabbak a chilei adatok, csak épp közülük azok esnek nagy súllyal latba, amik jobbak, míg a svédeknél azok, amik rosszabbak (mert azért a chilei 20 éves mortalitás még mindig jobb, mint a svéd 70 éves – ezen múlik a dolog)

A magyarázat szemléltetve



A standardizálás alapötlete

- Először is vegyük észre, hogy amikor korspecifikus rátákat használtunk, akkor lényegében egy confounding-ot oldottunk meg rétegzéssel
- Most egy módszert fogunk látni, mely a rétegzésnek legalább azt a problémáját megoldja, hogy nagyon sok eredményt kapunk
- (Persze a dolognak ára lesz!)
- A gond tehát az, hogy a korspecifikus mortalitásokat *különböző* korfákkal súlyozzuk
- Elég kézenfekvő ötlet: súlyozzuk *mindkét* országot *ugyanazzal* a korfával!

A standardizálás alapötlete

Korcsoport	Korspecifikus mortalitás		Létszám megoszlás (korfa)		
	Svédország	Chile	Svédország	Chile	WHO standard
0	0,0004	0,0010	0,1071	0,1523	0,1754
10	0,0002	0,0004	0,1324	0,1778	0,1706
20	0,0005	0,0008	0,1185	0,1544	0,1614
30	0,0006	0,0012	0,1403	0,1558	0,1475
40	0,0015	0,0023	0,1334	0,1435	0,1263
50	0,0041	0,0054	0,1361	0,0964	0,0992
60	0,0105	0,0138	0,1052	0,0635	0,0668
70	0,0290	0,0360	0,0736	0,0389	0,0373
80	0,0885	0,0909	0,0454	0,0147	0,0135
90	0,2386	0,1849	0,0079	0,0028	0,0019
100	0,5188	0,4440	0,0001	0,0001	0,0000

A svéd számolás akkor, ha chilei korfát használunk:

$$0{,}1523 \cdot 0{,}0004 + 0{,}1778 \cdot 0{,}0002 + \ldots + 0{,}0001 \cdot 0{,}5180 = 0{,}04659$$

Így már tényleg kisebb, mint a chilei érték!

A standardizálás alapötlete

Korcsoport	Korspecifikus mortalitás		Létszám megoszlás (korfa)		
	Svédország	Chile	Svédország	Chile	WHO standard
0	0,0004	0,0010	0,1071	0,1523	0,1754
10	0,0002	0,0004	0,1324	0,1778	0,1706
20	0,0005	0,0008	0,1185	0,1544	0,1614
30	0,0006	0,0012	0,1403	0,1558	0,1475
40	0,0015	0,0023	0,1334	0,1435	0,1263
50	0,0041	0,0054	0,1361	0,0964	0,0992
60	0,0105	0,0138	0,1052	0,0635	0,0668
70	0,0290	0,0360	0,0736	0,0389	0,0373
80	0,0885	0,0909	0,0454	0,0147	0,0135
90	0,2386	0,1849	0,0079	0,0028	0,0019
100	0,5188	0,4440	0,0001	0,0001	0,0000

A chilei számolás akkor, ha svéd korfát használunk:

$$0{,}1071 \cdot 0{,}0010 + 0{,}1324 \cdot 0{,}0004 + \ldots + 0{,}0001 \cdot 0{,}4440 = 0{,}01123$$

Így már tényleg nagyobb, mint a svéd érték!

A standardizálás alapötlete

- Hogy most az a svéd korfa, a chilei korfa, magyar korfa, vagy valami más, az első körben nem fontos kérdés, a lényeg, hogy *ugyanaz* legyen a korfa
- A kapott eredmény egy fiktív halálozási ráta lesz…
- …viszont összevethető a két populáció között!
- Értelme: mi *lenne* egy országban a halálozás, ha a korösszetétele olyan *lenne* mint a megadott korfa (de közben a korspecifikus mortalitások ugyanazok maradnának)
- Ha mindkettőnél ugyanazt adjuk meg, akkor értelemszerű, hogy eltüntettük az eltérő korösszetétel hatását

- Lényegében két lépést végzünk: rétegzünk, majd újra összerakunk – ahogy a rossz számítás is elmondható lenne, csak épp most az utóbbi lépésben kiküszöböljük azt, ami a rossz számításnál a problémát okozta
- A számérték függ a választott közös korfától, és akár a sorrendet is befolyásolhatja (de: ha az egyik korspecifikus görbe végig a másik felett húzódik, akkor lehetetlen, hogy megforduljon a sorrend, bármilyen korfát is választunk)

Referencia (vagy standard) populáció

- A fenti módszerrel mindig csak két országot tudunk összehasonlítani
- Gyakorlati okokból célszerű, ha lehetőleg mindenki ugyanahhoz a korfához standardizál, hiszen így az eredmények egymással is összevethetőek lesznek
- Nem csak adott párokat vizsgálhatunk, hanem egyszerűen minden ország közli a sajátját, de az egyezményes korfához standardizálva, így elég minden országnak egyetlen számot megadni, és mégis bármelyik ország bármelyikkel összevethetővé válik
- Ráadásul így kevésbé lehet játszani azzal, hogy olyan korfát választunk, amivel az jön ki, amit látni szeretnénk
- Ezért van néhány, nemzetközileg elfogadott ún. referencia, vagy standard populáció, melyeket általában használnak: pl. US Standard, Segi, ESP, WHO
- Ha standardizált eredményt közlünk, akkor odaírva, hogy mihez standardizáltuk, azonnal összevethető lesz az eredmény

A standardizálás menete

Korcsoport	Korspecifikus mortalitás		Létszám megoszlás (korfa)		
	Svédország	Chile	Svédország	Chile	WHO standard
0	0,0004	0,0010	0,1071	0,1523	0,1754
10	0,0002	0,0004	0,1324	0,1778	0,1706
20	0,0005	0,0008	0,1185	0,1544	0,1614
30	0,0006	0,0012	0,1403	0,1558	0,1475
40	0,0015	0,0023	0,1334	0,1435	0,1263
50	0,0041	0,0054	0,1361	0,0964	0,0992
60	0,0105	0,0138	0,1052	0,0635	0,0668
70	0,0290	0,0360	0,0736	0,0389	0,0373
80	0,0885	0,0909	0,0454	0,0147	0,0135
90	0,2386	0,1849	0,0079	0,0028	0,0019
100	0,5188	0,4440	0,0001	0,0001	0,0000

$$\mathrm{CDR}_{\text{Svédország}} = 0{,}1071 \cdot 0{,}0004 + 0{,}1324 \cdot 0{,}0002 + \ldots + 0{,}0001 \cdot 0{,}5188 = 0{,}010178$$
$$\mathrm{CDR}_{\text{Chile}} = 0{,}1523 \cdot 0{,}0010 + 0{,}1778 \cdot 0{,}0004 + \ldots + 0{,}0001 \cdot 0{,}4440 = 0{,}005548$$
$$\mathrm{DSR}_{\text{Svédország}} = 0{,}1754 \cdot 0{,}0004 + 0{,}1706 \cdot 0{,}0002 + \ldots + 0{,}0000 \cdot 0{,}5188 = 0{,}004316$$
$$\mathrm{DSR}_{\text{Chile}} = 0{,}1754 \cdot 0{,}0010 + 0{,}1706 \cdot 0{,}0004 + \ldots + 0{,}0000 \cdot 0{,}4440 = 0{,}005251$$

Amit elértünk (és amit nem)

- Egyetlen számban kaptunk confounding-ra szűrt eredményt
- (A fenti példa esetében a számértéknek önmagában nincs semmilyen értelme, csak összehasonlításban értelmezhető)

- De az esetlegesen eltérő rétegspecifikus hatásokat elfedi (és ez esetben a standard-választás is számíthat) – bár ilyenkor semmilyen egyetlen számba sűrítő index nem lesz az igazi
- További probléma, hogy bár a végeredmény csak egyetlen szám, valójában megjelenik a sok réteg problémája:
  - Ha több millió réteg van, akkor a standardnak annyi számot kell tartalmaznia
  - (Széleskörű megállapodás pedig csak néhány tényezőre van: nem, életkor)
  - Ráadásul igazából a kis létszámú rétegek problémája is megmarad: a kapott standardizált számérték nagyon bizonytalan lesz

Alkalmazási területek

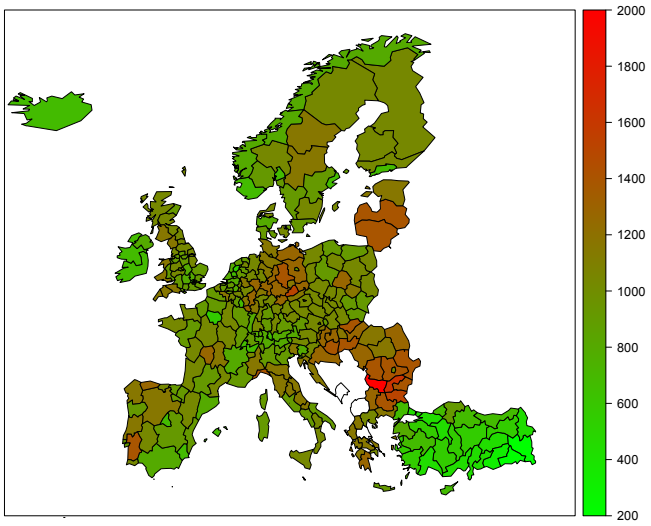
- Elvileg nem csak ilyen problémákra alkalmazható

	Nem szed HRT-t	Szed HRT-t
Alacsony szocioökonómiai státusz	4% (240/6000)	6% (12/200)
Magas szocioökonómiai státusz	1% (50/5000)	2% (60/3000)
Összességében	2,6% (290/11000)	2,3% (72/3200)

- Megtehetnénk például, hogy definiálunk egy „referencia szocioökonómiai státusz-fát” (pl. 55% magas, 45% alacsony) és ezzel súlyozva sűrítjük vissza egy számba a CV eseményeket
- Ekkor a szedők körében $6\% \cdot 0{,}45 + 2\% \cdot 0{,}55 = 3{,}8\%$ az arány, a nem szedők körében $4\% \cdot 0{,}45 + 1\% \cdot 0{,}55 = 2{,}35\%$
- Ez a fenti referenciára standardizált infarktus-rizikó (megoldódott a confounding!)
- Nagyon ritka, az alkalmazása lényegében a morbiditási/mortalitási adatokra korlátozott (erős történeti hagyományai vannak)

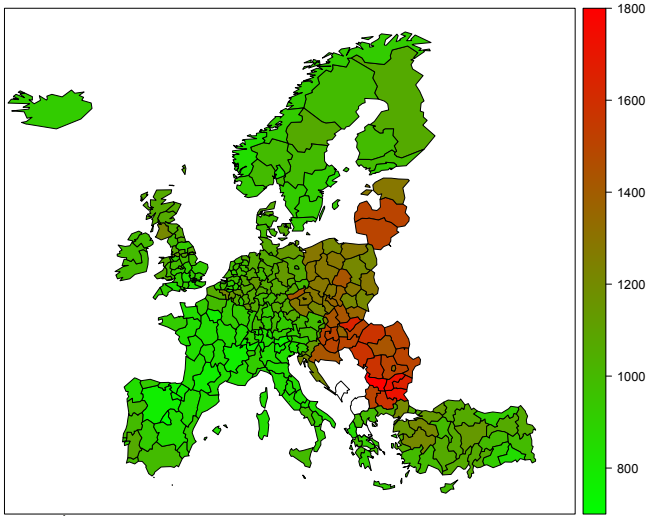
A standardizálás szemléltetése: térbeli különbségek

Nyers halálozási ráták:



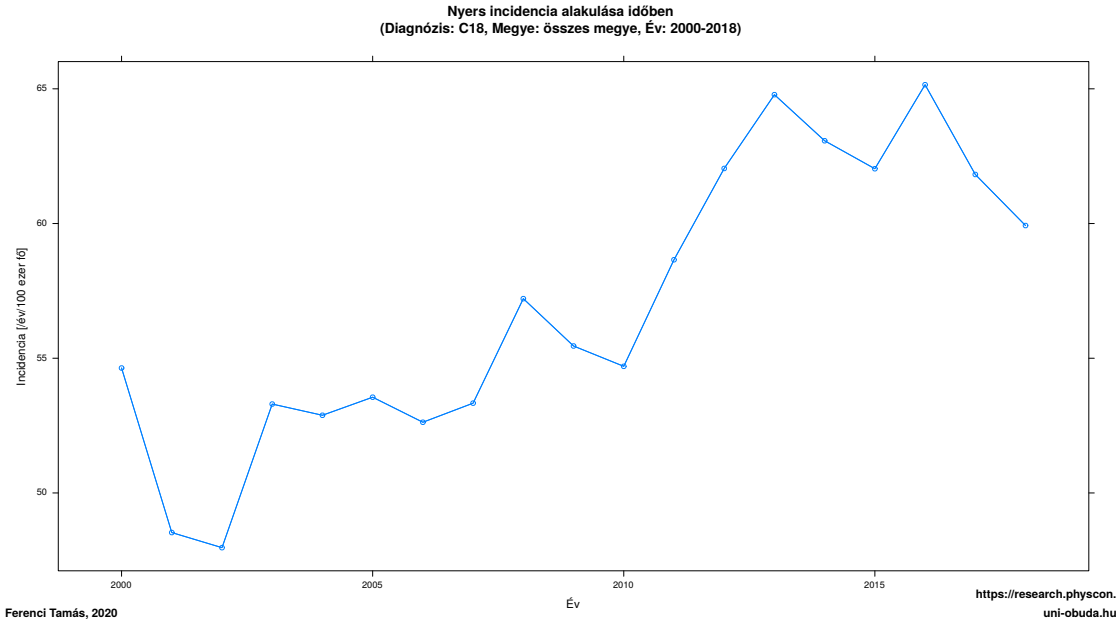
A standardizálás szemléltetése: térbeli különbségek

Standardizált halálozási ráták:



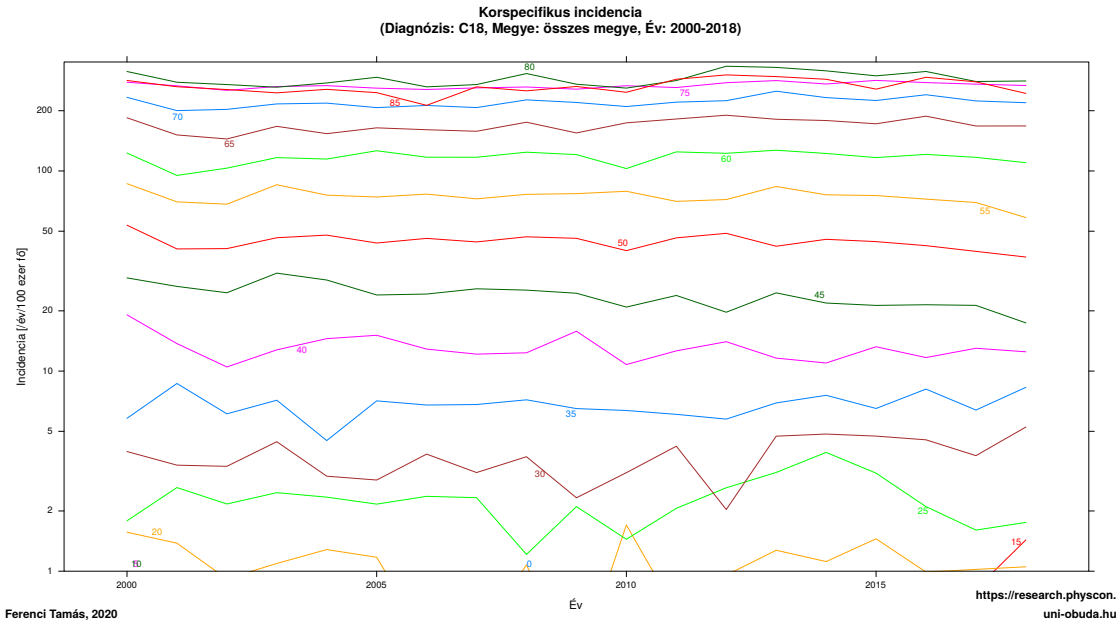
A standardizálás szemléltetése: időbeli különbségek

Nyers halálozási ráták:



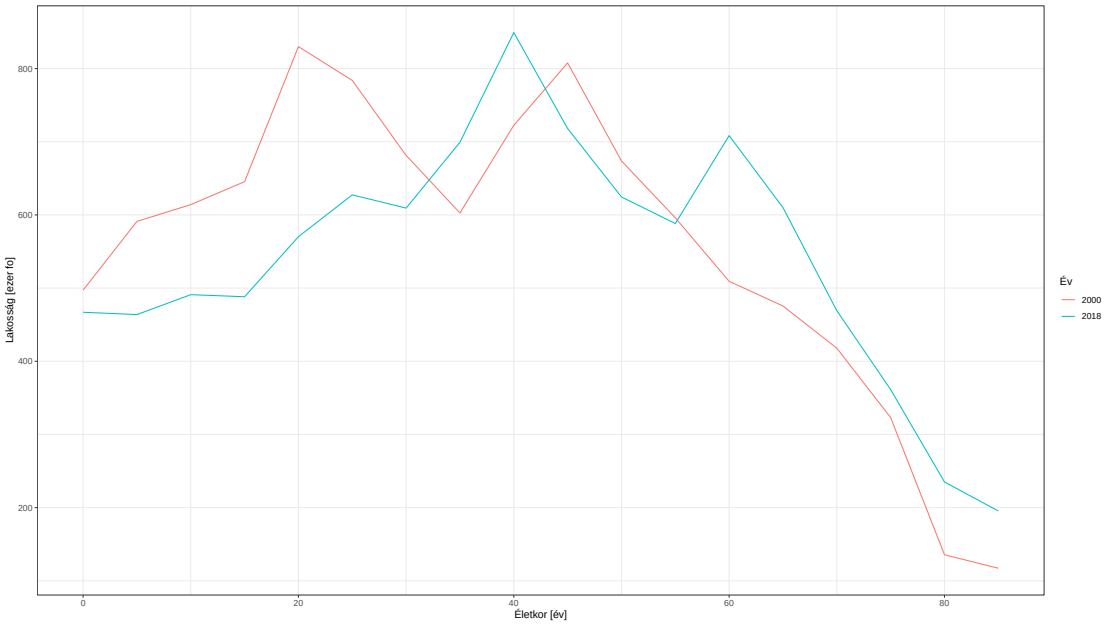
A standardizálás szemléltetése: időbeli különbségek

Korspecifikus incidenciák változása időben:



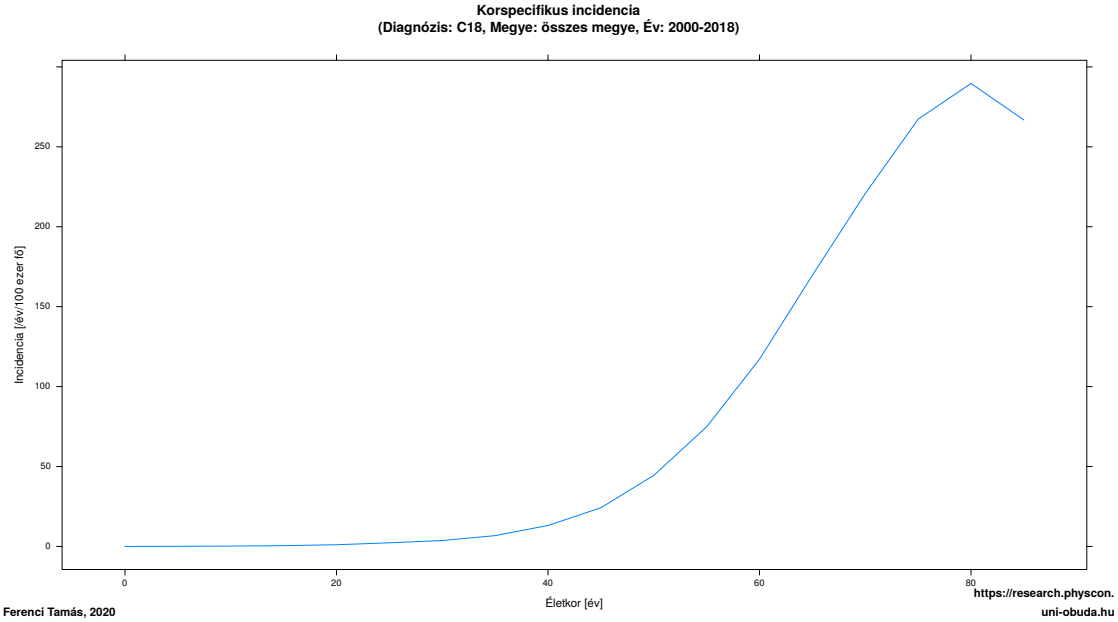
A standardizálás szemléltetése: időbeli különbségek

A társadalom öregedése:



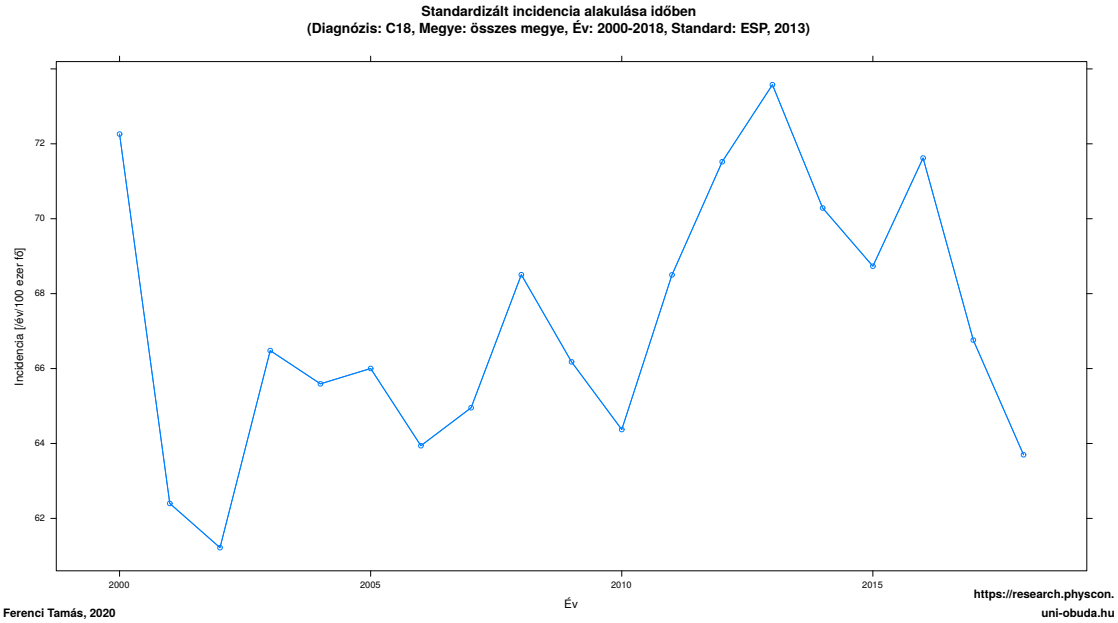
A standardizálás szemléltetése: időbeli különbségek

A rák korfüggése:



A standardizálás szemléltetése: időbeli különbségek

Standardizált halálozási ráták:



Illesztés (matching)

- Lényege: az összehasonlítandó csoportok alanyai között párokat képezünk, úgy, hogy azok tagjai lehetőleg minél hasonlóbbak legyenek, és így párosítva vizsgáljuk a végpontbeli eltéréseket, vagy más módon, de igyekszünk kiegyensúlyozni a csoportok confounderek szerint eltérő összetételét
- Előnyök:
  1. Ha nem is tökéletesen univerzális, de elég enyhe feltételek mellett nyújt jó választ, miközben általában a korábbiaknál több confounder is figyelembe vehető
- Hátrányok:
  1. Problémás lehet a „hasonlóság” meghatározása
  2. Minél több szempont szerint kell hasonlóságot számolni, tehát minél több confounder van, annál nehezebb/bizonytalanabb lesz a feladat (még ha nem is annyira, mint a korábbi módszereknél)
  3. Nem feltétlenül használja gazdaságosan fel a mintát
  4. Gond van, ha kettőnél több expozíció szerinti csoport van, amiket össze kell hasonlítani

Többváltozós – regressziós – modellezés

- Lényege: egy matematikai modellt tételezünk fel, melyben a végpontot mint változót valamilyen függvényforma szerint leírják a confounderek és az expozíció; ezt megbecsülve az expozíció hatása elkülöníthető (ceteris paribus értelmezés)
- Előnyök:

1. Jól képes számos confoundert felhasználni, melyek lehetnek folytonosak és kategoriálisak (akár vegyesen is), rendkívül flexibilis, ugyan sok feltevésre épít, de azok többsége jól ellenőrizhető

- Hátrányok:

1. Nagyon sok confounder itt sem kezelhető
2. A modellfeltevések teljesülése mindig kérdéses

Propensity score eljárások

- Lényege: meghatározzuk annak valószínűségét, hogy egy alany adott expozícióban részesüljön, a különböző ismert változói alapján, majd ezen ún. propensity score-ok alapján vagy hasonló párokat képezünk és körükben vizsgálódunk (PS matching), vagy ez alapján rétegzünk (PS stratification) vagy hagyományos elemzést végzünk, de úgy, hogy ez alapján meghatározott súlyokkal súlyozzuk az alanyokat (IPTW)

- Előnyök:

1. A többváltozós modellezéshez nagyon hasonló, talán kevesebb feltételezéssel a változók eloszlásáról és a köztük lévő kapcsolatokról (jelenleg is aktív vita tárgya)

- Hátrányok:

1. A regressziós modellhez nagyon hasonló (pontos összevetésük jelenleg is aktív vita tárgya)