# Az új statisztikai surveillance rendszer validációja

Ferenci Tamás, Kovács László, Surján György

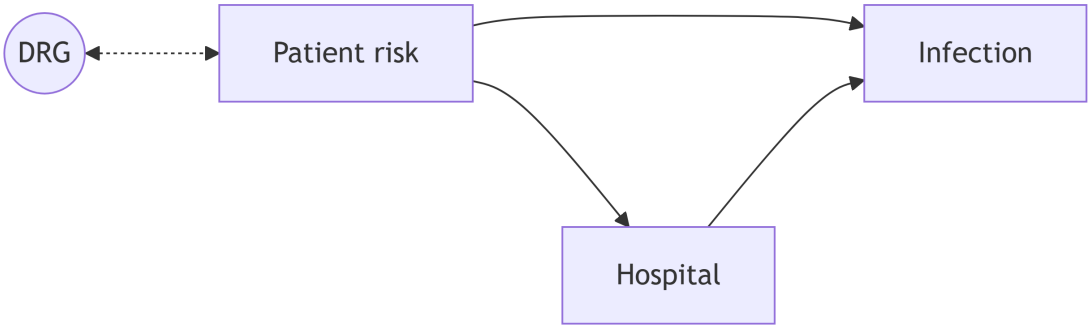
# Feladat

- A 2026 őszén elindított hazai nyilvános, intézmény-szintű adatközlési rendszer statisztikai módszertanának (SIR számolás) validálása

# Feladat

- A 2026 őszén elindított hazai nyilvános, intézmény-szintű adatközlési rendszer statisztikai módszertanának (SIR számolás) validálása
- Analitikus levezetés reménytelennek tűnik, a választott módszer: szimuláció

# Az egyszerű modell



# A problémák

A fenti modell nagyon könnyen szimulálható, ha mindhárom kategória egyetlen változónak felel meg (pláne, ha az folytonos), csakhogy most:

- A kockázat nem egyetlen, folytonos változó (nem, életkor, alapbetegség, súlyosság, társbetegségek stb.)

# A problémák

A fenti modell nagyon könnyen szimulálható, ha mindhárom kategória egyetlen változónak felel meg (pláne, ha az folytonos), csakhogy most:

- A kockázat nem egyetlen, folytonos változó (nem, életkor, alapbetegség, súlyosság, társbetegségek stb.)
- A proxy (HBCS) nem egyetlen, folytonos változó (több mint 700 kategória van)

# A problémák

A fenti modell nagyon könnyen szimulálható, ha mindhárom kategória egyetlen változónak felel meg (pláne, ha az folytonos), csakhogy most:

- A kockázat nem egyetlen, folytonos változó (nem, életkor, alapbetegség, súlyosság, társbetegségek stb.)
- A proxy (HBCS) nem egyetlen, folytonos változó (több mint 700 kategória van)
- A kórház nem egyetlen, folytonos változó (van majdnem 100 kórház)

# A megoldás alapvetései: a kórházi profil

- Az évek, a kórházak, a HBCS-ik, és az adott évben adott kórházban adott HBCS-ben ellátott betegek számát *nem* szimuláljuk újra, az le van rögzítve a magyar tényadatra

# A megoldás alapvetései: a kórházi profil

- Az évek, a kórházak, a HBCS-ik, és az adott évben adott kórházban adott HBCS-ben ellátott betegek számát *nem* szimuláljuk újra, az le van rögzítve a magyar tényadatra
- Amit szimulálunk, az *csak* a fertőzések száma (adott évben adott kórházban adott HBCS-ben)

# A megoldás alapvetései: a kórházi profil

- Az évek, a kórházak, a HBCS-ik, és az adott évben adott kórházban adott HBCS-ben ellátott betegek számát *nem* szimuláljuk újra, az le van rögzítve a magyar tényadatra
- Amit szimulálunk, az *csak* a fertőzések száma (adott évben adott kórházban adott HBCS-ben)
- Első fontos probléma: a szimulációban nem ismert a kórház profilja, valamint, hogy melyik HBCS mit jelent – tehát nem tudjuk követni azt, hogy adott kockázatú betegek hova kerülnek (ráadásul ezek eleve sem folytonos változók, hogy a "nagyobb kockázatú betegek a nagyobb kórházba kerülnek")

# A megoldás alapvetései: a kórházi profil

- Az évek, a kórházak, a HBCS-ik, és az adott évben adott kórházban adott HBCS-ben ellátott betegek számát *nem* szimuláljuk újra, az le van rögzítve a magyar tényadatra
- Amit szimulálunk, az *csak* a fertőzések száma (adott évben adott kórházban adott HBCS-ben)
- Első fontos probléma: a szimulációban nem ismert a kórház profilja, valamint, hogy melyik HBCS mit jelent – tehát nem tudjuk követni azt, hogy adott kockázatú betegek hova kerülnek (ráadásul ezek eleve sem folytonos változók, hogy a "nagyobb kockázatú betegek a nagyobb kórházba kerülnek")
- Alapgondolat: a "profil"-t a (szimulált) kockázatból határozzuk meg, lényegében visszafelé gondolkozva (erre mindjárt visszatérek)

# A megoldás alapvetései: a kórházi profil

- Az évek, a kórházak, a HBCS-ik, és az adott évben adott kórházban adott HBCS-ben ellátott betegek számát *nem* szimuláljuk újra, az le van rögzítve a magyar tényadatra
- Amit szimulálunk, az *csak* a fertőzések száma (adott évben adott kórházban adott HBCS-ben)
- Első fontos probléma: a szimulációban nem ismert a kórház profilja, valamint, hogy melyik HBCS mit jelent – tehát nem tudjuk követni azt, hogy adott kockázatú betegek hova kerülnek (ráadásul ezek eleve sem folytonos változók, hogy a "nagyobb kockázatú betegek a nagyobb kórházba kerülnek")
- Alapgondolat: a "profil"-t a (szimulált) kockázatból határozzuk meg, lényegében visszafelé gondolkozva (erre mindjárt visszatérek)
- Illetve valójában a kockázat sem ismert, tehát igazából a HBCS-ből számolunk

# A megoldás alapvetései: a kórházi profil

- Az évek, a kórházak, a HBCS-ik, és az adott évben adott kórházban adott HBCS-ben ellátott betegek számát *nem* szimuláljuk újra, az le van rögzítve a magyar tényadatra
- Amit szimulálunk, az *csak* a fertőzések száma (adott évben adott kórházban adott HBCS-ben)
- Első fontos probléma: a szimulációban nem ismert a kórház profilja, valamint, hogy melyik HBCS mit jelent – tehát nem tudjuk követni azt, hogy adott kockázatú betegek hova kerülnek (ráadásul ezek eleve sem folytonos változók, hogy a "nagyobb kockázatú betegek a nagyobb kórházba kerülnek")
- Alapgondolat: a "profil"-t a (szimulált) kockázatból határozzuk meg, lényegében visszafelé gondolkozva (erre mindjárt visszatérek)
- Illetve valójában a kockázat sem ismert, tehát igazából a HBCS-ből számolunk
- Mint a HBCS-beli kockázatok esetszámmal súlyozott átlaga

# A megoldás alapvetései: a kórház-hatás

- Ez a kórházi profil aztán hat a fertőzések számára

# A megoldás alapvetései: a kórház-hatás

- Ez a kórházi profil aztán hat a fertőzések számára
- Ez bizonyos szempontból egy fordítás: a valóságban a kórház profilja adott, és abból következik, hogy milyen kockázatú betegeket lát el, itt a kockázatot szimuláljuk, és abból határozzuk meg a “profil”-t

# A megoldás alapvetései: a kórház-hatás

- Ez a kórházi profil aztán hat a fertőzések számára
- Ez bizonyos szempontból egy fordítás: a valóságban a kórház profilja adott, és abból következik, hogy milyen kockázatú betegeket lát el, itt a kockázatot szimuláljuk, és abból határozzuk meg a “profil”-t
- (Más lehetőség nincs, hiszen a kórház valódi profilja ismeretlen a szimuláció számára)

# A megoldás alapvetései: a kórház-hatás

- Ez a kórházi profil aztán hat a fertőzések számára
- Ez bizonyos szempontból egy fordítás: a valóságban a kórház profilja adott, és abból következik, hogy milyen kockázatú betegeket lát el, itt a kockázatot szimuláljuk, és abból határozzuk meg a “profil”-t
- (Más lehetőség nincs, hiszen a kórház valódi profilja ismeretlen a szimuláció számára)
- A fordítás ugyan a kauzális viszonyokat a feje tetejére állítja, de ez pusztán adatgeneráláshoz irreleváns

# A megoldás alapvetései: a proxy-tökéletlenség

- A HBCS egy tökéletlen proxy-ja kell legyen a kockázatnak

# A megoldás alapvetései: a proxy-tökéletlenség

- A HBCS egy tökéletlen proxy-ja kell legyen a kockázatnak
- Eddig ezzel egyáltalán nem törődtünk (sőt, igazából a kockázat meg sem jelent, csak a HBCS)

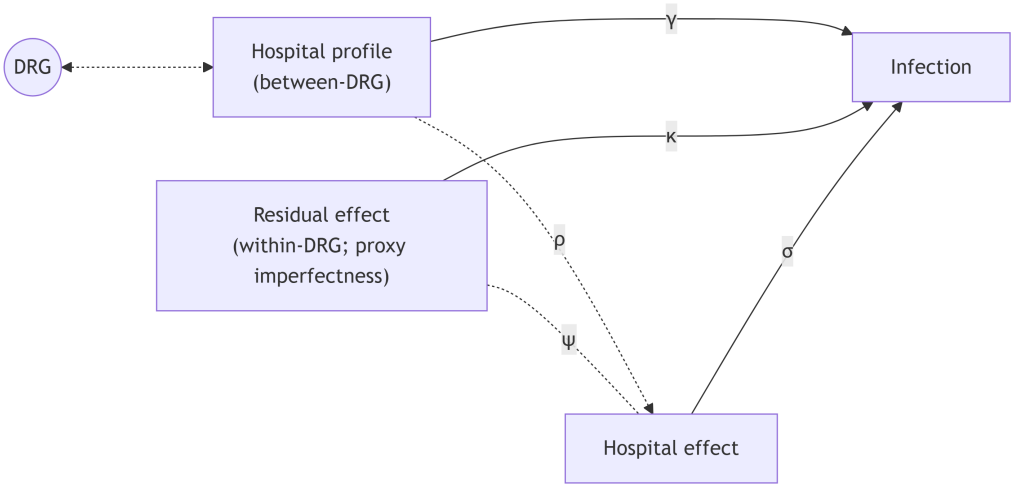
# A megoldás alapvetései: a proxy-tökéletlenség

- A HBCS egy tökéletlen proxy-ja kell legyen a kockázatnak
- Eddig ezzel egyáltalán nem törődtünk (sőt, igazából a kockázat meg sem jelent, csak a HBCS)
- Ezt úgy oldjuk meg, hogy felveszük egy vele – valamennyire, de nem tökéletesen – korrelált változót, ami *szintén* hat a fertőzések előfordulására

# A megoldás alapvetései: a proxy-tökéletlenség

- A HBCS egy tökéletlen proxy-ja kell legyen a kockázatnak
- Eddig ezzel egyáltalán nem törődtünk (sőt, igazából a kockázat meg sem jelent, csak a HBCS)
- Ezt úgy oldjuk meg, hogy felveszük egy vele – valamennyire, de nem tökéletesen – korrelált változót, ami *szintén* hat a fertőzések előfordulására
- Ez a maradék rész nem távolítható el (mert nincs rá proxy)

# A pontosabb modell



# A kód

```
DRGEffect <- rnorm(n_DRG)[drg_idx]

HospitalProfile <- collapse::fsum(DRGEffect * RD_CASE, g_hosp_all,
                                  use.g.names = FALSE) / caseByHosp
HospitalProfile <- (HospitalProfile - mean(HospitalProfile)) /
  sd(HospitalProfile)

zHE <- rho * HospitalProfile + sqrho * rnorm(length(HospitalProfile))

ProxyResid <- psi * zHE + sqpsi * rnorm(length(zHE))

HospitalEffect <- exp(sigma * zHE - sigma^2 / 2)

YearEffect <- runif(length(years), 0.8, 1.2)
```

# Pár magyarázat

- A “profil” standardizált

# Pár magyarázat

- A “profil” standardizált
- Kórház-hatás: a profilt “zajosítjuk”, precízen, `rho` erősséggel korrelált a szakmai profillal (ezt tudja megragadni a proxy!)

# Pár magyarázat

- A “profil” standardizált
- Kórház-hatás: a profilt “zajosítjuk”, precízen, `rho` erősséggel korrelált a szakmai profillal (ezt tudja megragadni a proxy!)
- Azért ilyen a megoldás, hogy a kórház-hatás varianciája 1 maradjon (továbbra is 0 átlag mellett), ez később lesz fontos, illetve a kórház-hatás és a profil korrelációja `rho` marad (`rho` változtatása tényleg csak a korrelációt befolyásolja)

# Pár magyarázat

- A “profil” standardizált
- Kórház-hatás: a profilt “zajosítjuk”, precízen, `rho` erősséggel korrelált a szakmai profillal (ezt tudja megragadni a proxy!)
- Azért ilyen a megoldás, hogy a kórház-hatás varianciája 1 maradjon (továbbra is 0 átlag mellett), ez később lesz fontos, illetve a kórház-hatás és a profil korrelációja `rho` marad (`rho` változtatása tényleg csak a korrelációt befolyásolja)
- A kórház-hatás `psi` erősséggel korrelált a maradék-résszel (ez az, ami nem távolítható el)

# Pár magyarázat

- A “profil” standardizált
- Kórház-hatás: a profilt “zajosítjuk”, precízen, `rho` erősséggel korrelált a szakmai profillal (ezt tudja megragadni a proxy!)
- Azért ilyen a megoldás, hogy a kórház-hatás varianciája 1 maradjon (továbbra is 0 átlag mellett), ez később lesz fontos, illetve a kórház-hatás és a profil korrelációja `rho` marad (`rho` változtatása tényleg csak a korrelációt befolyásolja)
- A kórház-hatás `psi` erősséggel korrelált a maradék-résszel (ez az, ami nem távolítható el)
- Hasonlóképp, a reziduális rész standardizált, és `psi` a korrelációja a kórház-hatással

# Pár magyarázat

- A “profil” standardizált
- Kórház-hatás: a profilt “zajosítjuk”, precízen, `rho` erősséggel korrelált a szakmai profillal (ezt tudja megragadni a proxy!)
- Azért ilyen a megoldás, hogy a kórház-hatás varianciája 1 maradjon (továbbra is 0 átlag mellett), ez később lesz fontos, illetve a kórház-hatás és a profil korrelációja `rho` marad (`rho` változtatása tényleg csak a korrelációt befolyásolja)
- A kórház-hatás `psi` erősséggel korrelált a maradék-résszel (ez az, ami nem távolítható el)
- Hasonlóképp, a reziduális rész standardizált, és `psi` a korrelációja a kórház-hatással
- A levonás azért van, hogy a várható érték ne menjen odébb (ezek ugye lognormális eloszlásúak)

# Zárólépés

- Mindhárom említett tényező (kórház-hatás, `sigma`, profil, `gamma`, reziduális rész, `kappa`) hat a fertőzés-számra

# Zárólépés

- Mindhárom említett tényező (kórház-hatás, `sigma`, profil, `gamma`, reziduális rész, `kappa`) hat a fertőzés-számra
- Sőt, még az év-hatás is

# Zárólépés

- Mindhárom említett tényező (kórház-hatás, `sigma`, profil, `gamma`, reziduális rész, `kappa`) hat a fertőzés-számra
- Sőt, még az év-hatás is
- Van még egy alap fertőzési ráta is, `r`

# Zárólépés

- Mindhárom említett tényező (kórház-hatás, `sigma`, profil, `gamma`, reziduális rész, `kappa`) hat a fertőzés-számra
- Sőt, még az év-hatás is
- Van még egy alap fertőzési ráta is, `r`
- (Összekapcsolás jósága nem kell, mert az is szorzó, vehetjük úgy, hogy be van építve az `r`-be)

# A kód

```
prob <-
  r *
  HospitalEffect[hosp_idx] *
  exp(gamma * DRGEffect - gamma^2 / 2) *
  exp(kappa * ProxyResid[hosp_idx] - kappa^2 / 2) *
  YearEffect[yr_idx] *
  p
prob <- pmin(prob, 0.99)

infection <- rbinom(n_RD, RD_CASE, prob)
```

# A szimuláció

- A `kappa` és a `gamma` rögzített (ezek hatása úgyis mérhető `psi` és `rho` segítségével)

# A szimuláció

- A `kappa` és a `gamma` rögzített (ezek hatása úgyis mérhető `psi` és `rho` segítségével)
- A többin végig-iterálunk grid-szerűen

# A szimuláció

- A `kappa` és a `gamma` rögzített (ezek hatása úgyis mérhető `psi` és `rho` segítségével)
- A többin végig-iterálunk grid-szerűen
- Az eredményeket lásd a preprint-ben!