

Lehen mahai-ingurua:
Ezagutza sortzen: frogetan oinarritutako praktikatik,
praktikak sortutako frogetara

3.3. Pazienteetatik datu-basera: informazioaren ibilaldi zoragarria

Ines Garmendia Navarro

BIOEF

ines.garmendianavarro@osakidetza.eus

a) Sarrera

1950eko hamarkadan, Isaac Asimov zientzia-fikziozko idazleak Multivac izeneko ordenagailu erraldoia aipatu zuen hainbat kontakizunetan. Izena egiazko ordenagailuen izenetatik hartu zen: ENIAC (1946) eta UNIVAC (1951), Estatu Batuetan fabrikatutako lehen ordenagailuak. UNIVAC hauteskunde presidentzial batzuen emaitza lehenengo aldiz (1952an) iragarri zuen ordenagailua izan zen (Eisenhower eta Adlai Stevenson-en artean, 100-1 lehenengoaren alde¹).

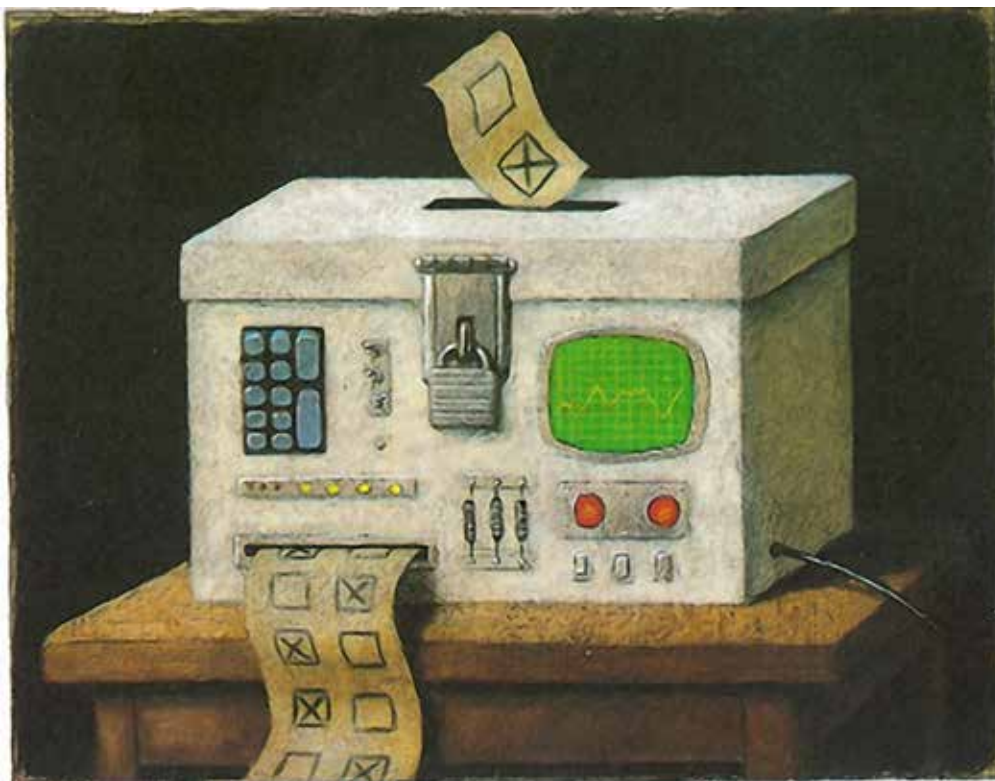


1. Irudia: UNIVAC, hauteskunde presidentzialen emaitza aurrean zuten makina (1952).

Jatorria: <https://arstechnica.com/tech-policy/2011/09/univac-the-troubled-life-of-americas-first-computer/>

¹ UNIVAC-en iragarpenak hain nabarmen egin zuten baten alde, non New York-eko CBS albiste kateen nagusiak, Mickelsonek, pentsatu zuen konputagailuaren errore bat izan behar zuela, eta erabaki zuen ez argitaratzea.

Multivac Asimoven asmakizuna zen, noski, baina zehaztasun handiz azaltzen zituen idazlearen ustean etorkizuneko ordenagailuek izango zituzten gaitasunak. Esate baterako, “Munduko arazo guztiak” (1959) kontakizunean, halaxe deskribatzen da Multivac-en funtzionamendua, Ben Manners protagonista, makinari galdera bat egiteko asmoz, haren aurrean esertzen denean:



2. Irudia. “Robbie eta beste kontakizun batzuk” liburutik hartuta / Isaac Asimov; hitzaurrea, Paul Cornell; oharrak eta jarduerak, María Raich, Juan Sánchez-Enciso; itzulpena, Koro Navarro; irudiak, David Shannon. Barcelona: Vicens-Vives, 2015.

“Ben Manners 5-B kabinan eseri zen, eta mantso-mantso tekleatu zuen [bere galdera] ... Ben erantzunaren zain geratu zen [...]. Oso denbora gutxian, galdera hura trilioi bat daturekin nahastu eta osotasun batean koordinatuko zen, eta Multivacek erantzunik egokiena aterako zuen osotasun hartatik.

Makinak “klik” egin zuen, eta txartel bat atera zen. Erantzun bat zekarren [...].”

Multivac ideal bat da, noski, baina gaur egun bizi dugun garaian -“Big Data”ren garaian- denok daukagu buruan, nola edo hala, ideal hori.

Multivacek bere baitan dauzkan datu guztiak nahasten ditu herritarrek egiten dizkioten galderari erantzuteko. Osasungintzaren arlora ere eman dezakegu ideia hori, argia baita horrelako makina batek izango lukeen abantaila: paziente bakoitzari buruzko galderak egiteko eta erantzuteko aukera izatea, klinikoan eta pazienteen eskura.

b) Saiakuntza klinikoak vs. eguneroko praktikaren ondorioz sortutako datuak erabiltzen dituzten ikerketak

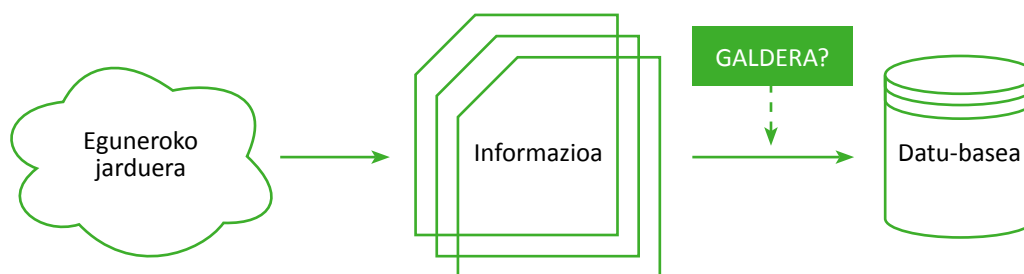
Oro har, ikerkuntza klinikoak galdera batekin hasi ohi dira. Mutur batean, “*gold standard*” diren saiakuntza klinikoak daude, arreta handiz eta prozedura zorrotzak jarraituz alde aurretik finkatutako galdera zehatz bati erantzuteko diseinatzen direnak. Ikerketa horietan, datuen bilketa galdera zehatz bati erantzuteko diseinatzen da, eta emaitzak metodo estatistiko jakin eta sendoak baliatuz aztertzen dira galderari modu zientifiko batez erantzuteko.

Hitzaldiak: lehen mahai-ingurua

Beste muturrean, osasungintzaren praktikaren beraren ondorioz jasotzen diren datuetan oinarritutako ikerketak dauzkagu. Azken urteotako aurrerapen teknologikoei esker, gaur egun milioka datu jaso, bildu eta prozesatzen dira egunero, osasungintzako aktibitatea xehetasun handiz jasotzen dutenak. Osasun-datuen pilaketa horri erantzunez, datuen “bigarren erabilera” egiten duten ikerketak sortu dira, berezko datu-planifikazio eza dela-eta erronka metodologiko berriak planteatzen dituztenak. Zeintzuk dira erronka horiek? Segituan ikusiko ditugu.

c) Mundu errealeko datuak erabiltzean sortzen diren erronkak

Esan dugun bezala, osasungintzaren eguneroko praktikan sortutako datuak erabiltzen dituzten ikerketetan, datuak aldeztu aurretik planifikatu gabe, sistemaren funtzionamenduaren beraren ondorioz sortzen dira. Ikerketa hauetan ere, ohikoa da galdera bati (edo gehiagori) erantzun nahi izatea. Horretarako, eskura daukagun informazio mordoa datu-base batean antolatu behar izaten da, metodo estatistikoak edo grafikoak aplikatu ahal izateko (ikus 3. irudia). Datu-base horien eraikuntzan sortzen diren arazo batzuk azalduko ditugu orain.



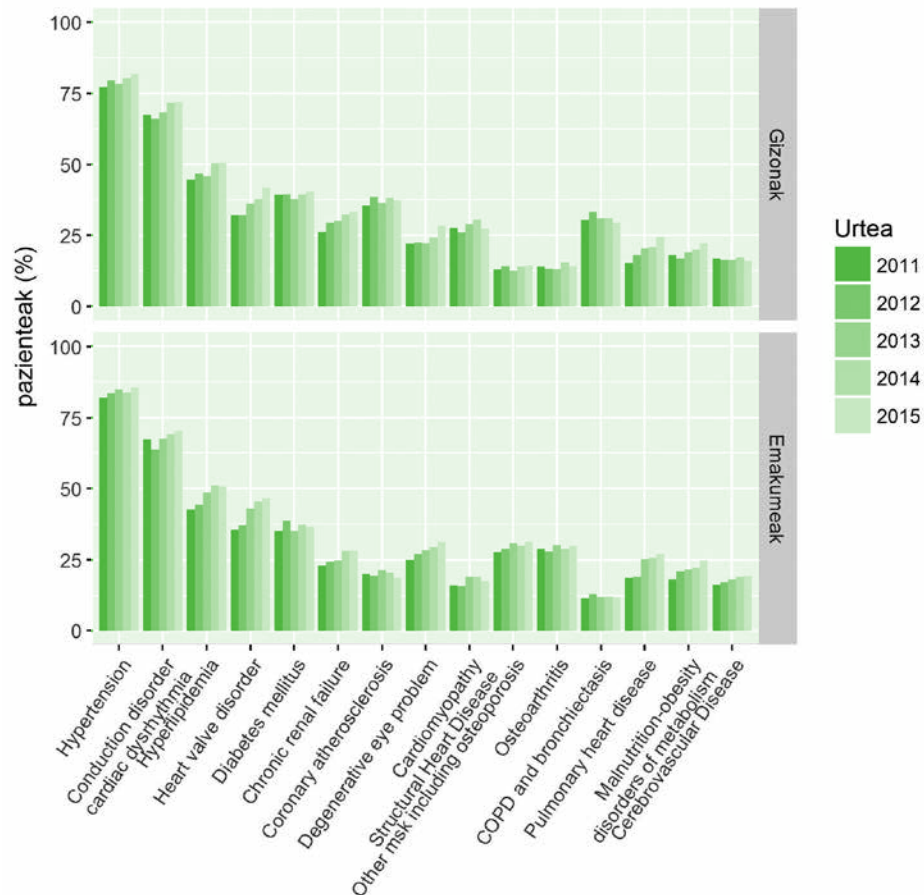
3. **Irudia:** Ikerketaren galdera nagusiak, edo helburuak, datu-basearen eraikuntza bideratuko du.

Datuak fidagarriak dira? Erroreak dituzte?

Neurketa-prozesuak berez zaratatsuak dira, datuak akatsekin grabatu daitezke eta datuen zati bat falta daiteke. Ikus dezagun adibide erreal bat: Lehen mailako arretako medikuek pazienteen tamaina eta altuera neurtu behar izaten dute sarritan, eta informazio hori sistema informatikoan grabatu ohi dute. Datu-base horietan, askotan, 89 eta 8900 bezalako balioak agertzen dira, unitate edo eskala ezberdinetan neurtu direla adierazten dutenak (bigarrena, 89 kilogramo=89000 gramotan). Kasu horietan datu “faltsurik” ez badago ere (hots, 8900 egiazko datu bat da), datu-basea inkoherentea izango da unitate-nahasketaren ondorioz. Beste batzuetan, datuak faltsuak edo inkoherenteak ez badira ere, horien interpretazioa da, edo izan daiteke, problematikoak. Ikus dezagun adibide bat.

Esanahia aldatu da? (Denboraren eragina)

Hainbat urtetako datuak aztertzen ditugunean, denboran zehar praktika bera aldatu ahal izan dela kontuan hartu beharko dugu. Behoko grafikoan (4. irudia) 2011-2015 urte bitartean bihotz-gutxiegitasunaren ondorioz Osakidetzako ospitaleetan ingresatutako pazienteak irudikatzen dira. Bost urte horietarako, bihotz-gutxiegitasunaz gainera pazienteek altaren momentuan zeuzkaten beste hainbat patologia bildu eta irudikatu dira (ardatz horizontalean, patologia ezberdinak; bertikalean, patologia horiek zeuzkaten paziente kopuruak, %etan emanak)



4. Irudia: 2011-2015 urte bitartean lehen aldiz bihotz-gutxiegitasunarengatik Osakidetzako ospitaleetan ingresatutako pazienteen patologia nagusiak, altaren uean.

Grafikoa lehenengo aldiz ikusten duenak pentsa lezake nabarmen areagotu dela azken urteetan pazienteen komorbiditatea (paziente gehiagok dituzte-eta ia patologia guztiak, 2011tik 2015era, grafikoak erakusten duen bezala). Ondorio hori, gainera, bat dator paziente hauen azken urteetako zahartzearekin (batez-besteko adina 79 urtekoa zen 2011an eta 81 urtekoa 2015ean): pazienteak zaharragoak direnean, patologia gehiago dituzte.

Baina, itxaron! Urte horietan zehar gaixotasunak hobeto kodifikatzeko ahaleginak egin ziren Osakidetzan, eta komorbiditatea areagotu dela egia bada ere (gaixo hauek zaharragoak direlako, besterik gabe), patologia hobeto kodifikatzeak ere eragina izan dezake grafikoa.

Datuak fidatzekoak badira ere, grafikoa kontu handiz interpretatu behar da.

Datuak koherenteak dira?

Jatorri ezberdinetako datuak gurutzatzen direnean, inkoherentziak sortu ohi dira. Ikus dezagun adibide bat. 4. irudian ikusi ditugun pazienteak bihotz-gutxiegitasunaren kohorte baten parte dira. Hantxe adierazi dugun bezala, 2011tik aurrera Osakidetzako ospitaleetan lehenengo aldiz ingresatutako pazienteei buruzko datuak bildu eta hainbat datu-iturri gurutzatu ditugu, OBI (Osakidetzako Business Intelligence) plataformako datu-baseak eta kanpoko erregistroak (Hilkortasun Erregistroa).

Datu-baseak aztertzerakoan zera ikusi dugu, inkoherentziak sortzen direla hilkortasun-datak eta ambulatorioko bisiten datak gurutzatzerakoan: paziente gutxi batzuek ambulatorioko "bisitak" dituzte hilkortasun-data pasa ondoren. Adituei galdetu eta gero, jakin izan dugu lehen arretako medikuek noizbehinka oharrak sartzen dituztela sistema informatikoan. Ez da benetako kontaktu bat, gestio administratibo bat baizik, eta, hortaz, ustezko inkoherentzia hori desagertu egiten da testuingurua ongi ulertu eta gero.

Hitzaldiak: lehen mahai-ingurua

Faltako balioak

Datu-falta da ikerkuntza klinikoan maiz gertatzen den arazo bat, askotan informazioa ez delako bere osotasunean jasotzen, zenbait aldagai ez direlako behar bezala jasotzen, edo pazienteek ikerketan parte hartzeari uzten diotelako. Mundu errealeko datuekin ere faltako balioak aurki daitezke (esate baterako, pazienteen analisiak egin eta grabatzen ez direnean, edo altuera edo neurria hartu baina zerbaitegatik behar bezala gordetzen ez direnean). Kasu horietan informazioa galdu egiten da, eta egoera hori jakina den neurrian, zenbateraino berreskuratu daitekeen ikusi beharko da.

Datu kopurua erraldoia denean, gainera, lan hori sistematizatu beharko da.

“Hutsuneak” nonahi!

Informazioa-falta ez denean ere, mota honetako datu-baseak eraikitzerakoan berezitasun bat sortu ohi da: paziente guztien datuak elkartu eta egituratzerakoan “hutsune” mordo sortu ohi da. Ikus dezagun adibide bat: askotan, metodo estatistikoak edo grafikoak erabili ahal izateko, patologia aldagai binario gisa kodifikatu behar izaten dira (hots, 1/0 balioak hartzen dituztenak, esate baterako, 1= pazienteak diabetesa badauka; 0= ez badauka). Horrela sortutako datu-basea 0-z beterik egongo da, eta teknika estatistiko proposak erabili beharko dira egoera horri aurre egiteko.

Informazio-eza informazioa da!

Aurrekoarekin lotuta, kasu batzuetan informazio-eza informazio gisa interpretatzea komeni da. Ikus dezagun adibide bat: argi dago Osakidetzako pazienteek medikura egiten dituzten bisita-kopuruen arabera aurkituko ditugula paziente horien datuak informazio-sistemetan. Esate baterako, ambulatortiora maiz joaten diren pazienteei (zaharrenak edo gaixotasun kronikoak dituztenak, beharbada) probabilitate handiagoz hartuko zaizkie neurriak, baina medikuarenera gutxi joaten direnak, aldiz (paziente osasuntsuak), seguru asko ez dute izango erregistro bakar bat ere informazio-sisteman.

Egoera horretan, “datu-ezak” berak informazioa emango digu -medikura ez joatea osasun onaren seinale den neurrian-.

Kontzeptu bakarra, aldaera ezberdinak (informazio ez-egituratua)

Orain arte informazio egituratua ikusi dugu (hots, egitura finkoarekin, tauletan gorde daitezkeen datuak). Baina osasungintzaren eguneroko praktikan sortzen den informazio gehiena testu edo irudi formatoan gordetzen da. Informazioa ez-egituratua denean, erronka berri eta latzak sortzen dira.

Klinikoek idatzitako testuetan, hainbat sinonimo edo esanahi bereko esaldi erabiltzen dira gauza berbera adierazteko. Kardiologian, esaterako, kanporatze-frakzioa hainbat modu desberdinetan azaltzen da idatzia, hala nola “fe”, “fracción de eyección” (gaztelaniaz), edo funtzio sistolikoaren aipamenaren pean.

Testu gutxi batzuk aztertzen ditugunean erraza da barianteak kategoria gutxi batzuetara eramatea, baina datu-kopuru erraldoiekin prozedura automatikoak eraiki eta modu sendo batean aplikatu beharko dira, bai testuak eta bai irudiak informazio bilakatu ahal izateko.

d) Zer soluzio daude?

Datu-base txiki bat daukagunean, goian aipatutako arazo gehienek konponbide erraza dute, iker-tzaileek banan-banan konpondu ahal izango dituzte-eta. Baina, esan dugun bezala, datu-kopuru erraldoiekin beharrezkoa da garbiketa-prozesu eta transformazioak sistematizatzea. Ba al dago lan hori egiteko gaitasuna izango duen makinarik?

“Frontiers in massive data analysis” liburuan argibide batzuk ematen dira datuak behar bezala pres-tatu eta etekin handiena atera ahal izateko. Zehazki, datuen garbiketa sistematizatzeko [2,100 orr.]:

- Datuak jaso eta biltzen dituen softwarean bertan auto-garbiketa mekanismoak sar litezke. Tresna horiekin batera, idealki, auditoria bat ezarri beharko da egindako aldaketak aztertu eta behar den kasuetan desegin ahal izateko;
- Zenbait arazo aldeztatik ezagunak badira -hala nola unitate edo eskala-nahasketak-, erre-gela bidez konpondu edo erresistenteak izango diren eredu estatistikoak eraiki beharko dira;
- Faltako balioena arazo metodologiko ezagun bat da, eta soluzio algoritmiko ugari eman izan dira: batzuetan, datuak osatu egiten dira ereduak kalkulatu baino lehen (datu-kopuru erral-doiaren eskalan, adibidez, praktikoa izan daiteke falta diren balioak batezbestekoaz osatzea); besteetan, algoritmoek bere kontuan hartzen dituzte hutsune horiek. Edozein kasutan, fun-tsezkoa da informazio faltaren zergatia ulertzea.
- Datu ez-egituratuak erabili ahal izateko (testuak eta irudiak), lengoia naturala prozesatzeko metodologiak menperatu eta erabili beharko dira, kasurik gehienak ondo konponduko dituen teknologia sendo bat sortu eta aplikatzeko.

Azkenik, datuak garbitu eta aurreprozesatzeko metodoa gorabehera, pausoak modu egokian doku-mentatu beharko dira eta, ahal den neurrian, erabilitako programaren *script*-a edota kodea datuekin batera gorde beharko dira. Datuak garbitzeko eta prestatzeko pausoak leialtasunez erreproduzitzeko eta pausoak berez ulertzeko gai izan beharko genuke.

e) Amaitzeko

Osasungintzaren eguneroko praktikan biltzen diren datu-baseekin sor daitezkeen hainbat arazo ikusi ditugu, zailtasun maila ezberdinak dituztenak, datu-kopurua erraldoia denean bereziki sortzen direnak.

Erroreak modu automatizatu batean detektatzea eta konpontzea oso zaila izan daiteke eta datu-kopuru handiak edukitzeak bakarrik ez digu arazo hori konponduko. Eta datuak errorerik gabekoak badira ere, interpretazioa problematikoa izan daiteke batzuetan. Azkenik, datu-baseak gurutzatzean, inkoherentziak sor daitezke.

Edozein eredu matematiko edo grafiko erabiltzerakoan, ondorio egokiak atera ahal izateko datuak errorerik gabekoak izan behar dira; eta, are garrantzitsuagoa beharbada, datuen jatorria eta testuin-gurua ondo ezagutu beharko da.

Asimovek 1950eko hamarkadan irudikatutako Multivac bat izango al dugu guk etorkizunean? Neurri handi batean, gure ahalmenaren araberakoa izango da erantzuna, hots, galdera bakoitzari erantzuten dioten datu-base egokiak eraikitzeak ahalmenaren araberakoa... Eta, nolana ere, galdera asko erantzunik gabe geratuko dira.

f) Erreferentziak

[1] Robbie : eta beste kontakizun batzuk / Isaac Asimov ; hitzaurrea, Paul Cornell ; oharrak eta jarduerak, María Raich, Juan Sánchez-Enciso ; itzulpena, Koro Navarro ; irudiak, David Shannon. Barcelona : Vicens-Vives, 2015

[2] National Research Council. (2013). *Frontiers in massive data analysis*. National Academies Press.

[3] Data, M. C. (2016). *Secondary Analysis of Electronic Health Records*. Springer International Publishing.