ANDMEKADU JA VEAD
NING NENDEGA KAASNEVATE TAKISTUSTE LAHENDAMINE

Viktoria Kirpu
Natalja Eigo

Tervise Arengu Instituut

2018

Märksõnad: kadu, vead, imputeerimine, lisainformatsioon, nihkega hinnang, nihketa hinnang

Sissejuhatus

Tervisestatistikat kogub, töötleb, analüüsib ja avaldab Eestis Tervise Arengu Instituut (TAI). Üheks statistika andmeallikaks kasutab TAI E-Tervise infosüsteemi ehk TIS-i. Kahjuks esineb vaadeldavas andmebaasis puudujääke andmete kaetuses ja kvaliteedis. TIS-i andmete valideerimiseks kasutab TAI Eesti Haigekassa saadetud raviarvete andmeid. Tänaseks on haigekassale laekunud rohkem vaatlusi kui tervise infosüsteemi.

Kui aga analüüsitavas andmebaasis on andmed jäetud saatmata, siis on oluline statistika tegemisel sellega arvestada ja rakendada vajalikke statistilisi meetodeid. Vastasel juhul tehakse puudulike andmete põhjal reaalsele olukorrale mittevastavad järeldused.

TIS andmed

2008. aastal loodud E-Tervise infosüsteem ehk TIS (Digilugu), mida haldab ja arendab nüüd Tervise ja Heaolu Infosüsteemide Keskus (TEHIK), on erinevaid lahendusi hõlmav tervishoiusektori koostöömudel, mille üheks oluliseks osaks on riigi infosüsteemi kuuluv andmekogu [3]. Sellega liidestunud tervishoiuasutused saadavad sinna haiguslugude kokkuvõtteid ehk epikriise ning teisi meditsiinidokumente. TIS-i andmeid kasutatakse muuhulgas tervislikku seisundit kajastavate registrite pidamiseks ja tervisestatistika tegemiseks. [11] Tervishoiuteenuse osutajatel on kohustus seaduse järgi edastada määratletud andmed tervise infosüsteemi [12].

Haigekassa andmed

Eesti Haigekassa tähtsaim ülesanne on korraldada ravikindlustust, et võimaldada kindlustatud isikutele ravikindlustushüvitisi. Lisaks on haigekassa ülesanne aidata kaasa ka ravistandardite ja ravijuhiste koostamisele, motiveerida tervishoiuasutusi arendama tervishoiuteenuste kvaliteeti, korraldada ravikindlustust ja haigekassat puudutavate välislepingute täitmist; osaleda tervishoiu planeerimisel; avaldada arvamust haigekassa ja ravikindlustusega seonduvate õigusaktide ja välislepingute eelnõude kohta ning anda nõu ravikindlustusega seonduvates küsimustes. [4] Lisaks sellele kogub haigekassa tervishoiuteenust osutavatelt asutustelt ravijuhtumite raviarvete kohta dokumentatsioone.

Andmete kaoga kaasnevad takistused

Andmekadu analüüsitavas andmebaasis tähendab olukorda, kui oli meditsiinitöötaja poolt registreeritud ravijuht, kuid selle kohta dokumentatsiooni Tervise Infosüsteemi (TIS) pole esitatud.

Andmete puudumise tõttu ei kao ainult vajaminev informatsioon ega vähene uuringu võimsus[1], vaid tekkivad nihkega hinnangud[2] [14]. Avastamata kadunud vaatluste mahtu ehk ravijuhude arvu, mille kohta dokumentatsiooni ei esitatud TIS-i ja kontrolli käigus nende puudumist ei avastatud, on tähtis minimiseerida. Vastasel korral statistilised järeldused, näiteks tulemuse usaldusintervall, on ilmselt tehtud valesti. [8] Kadudega mitte arvestamine suurendab saadud hinnangute nihet. Kvaliteetse statistika jaoks on vajalik omada nihketa hinnangut[3] või tuleks nihet muuta võimalikult väikseks. Mida väiksem on nihe,

[1] Indikaator (tõenäosus), mis hindab kui oluline saadud tulemus statistilises mõistes on [2].

[2] Hinnang, mille keskväärtus erineb hinnatava parameetri tõelisest väärtusest teatava süstemaatilise vea võrra [9].

[3] Hinnang, mille keskväärtus võrdub hinnatava parameetri tõelise väärtusega [9].

seda paremini peegeldavad statistilised tulemused reaalset olukorda.

Näiteks jääb tervise infosüsteemis arstide poolt suuremal määral esitamata ravijuhu erakorralisuse tüüp. Kui tekib olukord, kus arstid jätavad erakorraliste ravijuhtumite korral vaadeldava tunnuse märkimata, siis jääb mulje, et erakorralisi ravijuhtumeid esineb meie riigis vähe. Sellises olukorras saame olla kindlad, et saadud statistika ei kirjelda tegelikust ning oleme saanud nihkega hinnangud [8]. Samal põhimõttel tekivad nihked ka siis, kui mõned epikriisid on koguni jäänud dokumenteerimata.

Levinud on väär arusaam, et kui vastamismäär on kõrge, siis pole oluline arvestada andmete kaoga. Statistikas ei keskenduta vastamismäärale kui indikaatorile, mis vähendab kadudest põhjustatud nihet, sest iseenesest vastamismäär ei mõõda seda. [14] Erinevalt dispersioonist ei lähene nihe valimimahu suurenedes nullile [7, 8]. Kaost põhjustatud nihke vähendamiseks on oluline kasutada vajalikke hindamismeetodeid [14].

Kadudeta andmestiku käsitlemine

Proovime näiteks leida mingisuguse tunnuse Y kogusummat. Tähistagu $U=\{1,\ldots, k,\ldots, N\}$ üldkogumit suurusega N ning y_k tunnuse Y väärtus k-nda objekti korral. Sel juhul saame, et tunnuse Y kogusumma on:

$$Y = y_1 + \cdots + y_N = \sum_{k=1}^{N} y_k \ .[3]$$

Antud juhul saame teha statistikat kui TIS-i andmestik on täielik ehk sinna on saadetud kõik ravijuhude epikriisid ja on olemas kõikide vaatluste korral tunnuste väärtused [14].

Vigade ja kadudega andmestiku käsitlemine

Empiirilistes andmetes esineb peaaegu alati kadu. Puuduvate väärtustega andmed tekivad siis, kui valimisse sattunud vaatlusel ei ole andmeid täielikus mahus ehk kui toimub nn objekti kadu (ingl k *unit non-response*) või kui esineb osaliselt muutujate kadu ehk toimub nn väärtuse kadu (ingl k *item non-response*) [1, 6]. Antud juhul objekti kadu tähendab epikriisi mitteesitamist TIS-i ja väärtuse kadu kui epikriis on esitatud puudulike andmetega.

TIS-is esinevaid vigu saab jagada järgmiste tüüpidesse:

- **juhuslikud vead** (tingitud mõõtmise või protokollimise ebatäpsusest, üldreeglina muudavad tulemust vähe ja on raskesti avastatavad), näiteks patsiendi sünnikuu vale märkimine;
- **süstemaatilised vead** (tingitud enamasti instrumendi ebatäpsusest), näiteks kui arst kasutab kõikide haiguste kirjeldamiseks ühte ja sama diagnoosi;
- **jämedad vead** (tunnuse väärtus on väljaspool tunnuse võimalike väärtuste piirkonda), näiteks emakakaelavähi diagnoos meessoost patsiendil;
- **loogilised vead**, kus erinevate tunnuste väärtused ei ole kooskõlas, näiteks inimese haiglast väljakirjutamise kuupäev on märgitud pärast surmakuupäeva. [5]

Vigade avastamisel on vaja need elimineerida ja vajadusel käsitleda neid kui kadusid.

Kadudega andmestikku ei saa töödelda samamoodi nagu andmeid, kus kadu puudub. Peamised meetodid kaoga andmete käsitlemisel on **lisainformatsiooni kasutamine** ja **imputeerimine**[4]. [14]

Lisainformatsiooni kasutamine

Tervise Arengu Instituut kasutab TIS-i andmekao puhul andmete täiustamiseks lisaallikana haigekassa andmeid.

Kaoga andmestike korral eristatakse kolme tüüpi lisainformatsiooni: InfoU, InfoS, InfoUS [8, 13].

- **InfoU**
  Sellist tüüpi info korral on abiinformatsiooniks vektor $x_k = x_k^*$, mis on teada iga $k \in N$ korral. Lisainformatsioon, mis on kogusummade vektor $X^* = \sum_U x_k^*$, on teada üldkogumi U tasemel.
- **InfoS**
  Sellist tüüpi info korral on abiinformatsioon teada olemasolevate andmete piires, aga mitte üldkogumi tasemel ning abivektoriks $x_k = x_k^\circ$.
- **InfoUS**
  Sellist tüüpi info korral on teada üldkogumi kui ka valimi tasemel, $x_k = \begin{pmatrix} x_k^* \\ x_k^\circ \end{pmatrix}$. [14]

Lisaallikast leitud väärtused on vaadeldava objekti tegelikud andmed. Seega need andmed lisatakse käesolevasse andmestikku. Sellest tingituna on saadavad hinnangud nihketa.

TIS-i andmete „rikastamiseks“ püüab TAI kasutada lisainformatsioonina *InfoU* haigekassast saadud andmeid. TAI eeldab, et viimasena mainitud andmeallikas on info teada kogu üldkogumi tasemel.

Imputeerimine

Väärtuste imputeerimismeetodeid jagatakse kolme gruppi:

- statistilised prognoosimise meetodid;
- mittevastanutega sarnastelt vastanud objektidelt väärtuse omistamine;
- eksperthinnang [14].

Esimest kahte gruppi liigitatakse statistilisteks meetoditeks, sest nad kasutavad levinud meetodeid asendusväärtuste leidmiseks. Esimese grupi meetodid põhinevad eeldataval seosel tunnuste vahel nagu näiteks regressioonprognooside leidmisel. Teise grupi meetodeid võib nimetada doonoripõhisteks, sest imputeeritud väärtus laenatakse mõnelt vaadeldud objektilt, mis on sarnane puuduvale objektile. Kolmanda grupi meetodid sõltuvad palju eksperdi oskustest ja teadmistest. [14]

Imputeerimismeetodeid võib liigitada ka deterministlikeks, st imputeerimisprotseduuri korrates jõutakse alati täpselt samale tulemusele, või juhuslikeks, kui protseduuri korrates saadakse erinevad imputeeritud väärtused. Regressiooniimputeerimine prognooside abil on näide deterministlikust meetodist. Kui imputeerime aga juhuslikult valitud sarnase objekti

[4] protseduur, kus ühe või mitme tunnuse puuduvad väärtused täidetakse olemasolevate andmete või mõne muu informatsiooni põhjal [10].

väärtuse, siis on tegu juhusliku meetodiga, mida tihti kasutatakse *Hot-Deck* imputeerimiseks. [14]

Oluline on arvestada, et imputeeritud väärtused on kunstlikud –nad on kas konstrueeritud mingi reegli kohaselt või teiste vastanud objektide väärtused. Seetõttu erinevad imputeeritud väärtused alati mingil määral objektide tegelikest väärtustest. Oodatakse, et imputeerimise korral oleksid nad väikese hajuvusega ja nihketa või, et nihe oleks väike. [14]

Üldiselt tuleb imputeerimisse suhtuda ettevaatlikkusega, sest me püüame teha usaldusväärset statistikat andmetelt, mille kohta on juba alguses teada, et need on vähemal või rohkemal määral ebatäpsed. Samas on praktilistel põhjustel seda sageli vaja teha. Ei ole mõjuvat põhjendust selle kohta, et hoolikas imputeerimine tekitaks suuremat kahju hinnangutele kui teised meetodid statistika tootmisel. [14]

Kokkuvõte

Kokkuvõtteks võib öelda, et statistika tegemiseks on oluline teada, kas me kasutame andmete analüüsimiseks täielikku või kadudega andmestikku. Vastavalt sellele eelteadmisele rakendatakse statistika tegemiseks erinevaid meetodeid.

Kadudeta andmete korral saab kohe teha statistikat. Kadudega ja/või vigadega andmete puhul on tarvis teha eelnevat andmete töötlemist. Selleks kasutatakse erinevaid meetodeid, näiteks lisainformatsiooni kasutamist ja/või imputeerimist ning vastavalt sellele prognoositakse puuduolevad väärtused. Kui teha ekslikult statistikat kadudega andmetega, siis saadakse tulemusteks nihkega hinnangud, mis ei vasta tegelikkusele. Puuduolevatele andmetele väärtusi leides on võimalik nihet vähendada või parimal juhul isegi kaotada.

Seega tervisestatistika tegemiseks tuleb kontrollida TIS andmete kvaliteeti ehk kas tervise infosüsteemi on saadetud kõikide ravijuhtude vajalikud dokumentid.

Käesolevas TIS-i andmekvaliteedikontrollis on TAI-l soov kasutada lisainformatsioonina *InfoU* haigekassast saadud andmeid. Tänu sellele saaksime teada, kui palju andmeid jääb tervise infosüsteemi esitamata ning kui palju tehakse andmete esitamisel vigu. Vastavalt sellele oleks võimalik kasutada tervise infosüsteemi andmestiku täiendamiseks lisainformatsiooni ja imputeerimist.

Kasutatud allikad

[1] Andridge, R. R., Little , R. J. A. (2010). *A Review of Hot Deck Imputation for Survey Non-response*. Int Stat Rev. 2010 aprill, 78(1), 40–64. doi: 10.1111/j.1751-5823.2010.00103.x

[2] Aron, A., Aron, E. N. (1997). *Statistics for the behavioral and social sciences: A brief course*. Upper Saddle River, NJ: Prentice Hall.

[3] Eesti E-tervise Sihtasutuse kodulehekülg: *Tervise infosüsteem*. (n.d.). Kasutatud 06.02.2018. *http://www.e-tervis.ee/index.php/et/eesti-etervise-sihtasutus/tervise-infosusteem*

[4] Eesti Haigekassa kodulehekülg: *Haigekassa organisatsioon*. (n.d.). Kasutatud 06.02.2018. *https://www.haigekassa.ee/haigekassa/organisatsioon*

[5] Käärik, E. (2013). *Andmeanalüüs II*. Loengukonspekt. Kasutatud 06.04.2018. *http://dspace.ut.ee/bitstream/handle/10062/35401/AndmeanaluusII.pdf?sequence=1*

[6] Lepik, N., Traat, I. (2016). *Tõenäosuslik valikuuring I*. Loengukonspekt. Tartu Ülikool, matemaatilise statistika instituut. Kasutatud 08.02.2018. *https://courses.ms.ut.ee/MTMS.01.003/2016_fall/uploads/Main/loengud\\2016.pdf*

[7] Shouten, B. Cobben, F. (2007). *R-indexes for Comparison of Different Fieldwork Strategies and Data Collection Modules*. Discussion paper 07002. Voorburg-Heerlen: Statistics Netherland.

[8] Särndal, C.-E. Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley & Sons: New-York.

[9] Termeki: *Andmeanalüüsi ja statistika oskussõnastik* (2016). Kasutatud 30.11.2017. *https://term.eki.ee/termbase/view/8917007/et/en/?initial=N\#/concept/view/1718212121/*

[10] Termeki: *Statistiline imputeerimine*. (*n.d.*). Kasutatud 20.11.2017. *http://termin.eki.ee/esterm/concept.php?id=29907\&term=statistiline\%20imputeerimine*

[11] Tervise Arengu Instituudi kodulehekülg: *E-tervise infosüsteem*. 09.08.2017. Kasutatud 31.01.2018. *http://www.tai.ee/et/tegevused/tervisestatistika/tervise-infosusteemi-statistikamoodul*

[12] Tervishoiuteenuste korraldamise seadus1, § 592. (01.01.2018). Riigi Teataja I. Kasutatud 09.04.2018. *https://www.riigiteataja.ee/akt/TTKS*

[13] Toompere, K. (2006). *Kadu ja kao indeksid*. Bakalaureusetöö. Tartu Ülikool, matemaatilise statistika instituut.

[14] Toompere, K. (2009). *Imputeerimis- ja kaalumismeetodite mõju hinnangute nihkele*. Magistritöö. Tartu Ülikool, matemaatilise statistika instituut.