

Andmekaitsealane mõjuhindang projekti „Kriisihaavatavuse kujunemine, tulevikukriisid ja sekkumised ühiskondliku valmisoleku suurendamiseks“ raames haavatavusindeksi koostamiseks

Kati Orru, Tartu Ülikool

19.05.2026

Järgnev hinnang käsitleb uuringu raames kavandatud isikuandmete töötlemise eesmärgi, õiguslikke aluseid, vajalikkust ja proportsionaalsust, andmetöötlusega seotud võimalikke riske ning nende maandamiseks rakendatavaid kaitsemeetmeid kooskõlas isikuandmete kaitse üldmääruse (IKÜM) nõuetega.

1) Mis eesmärgi saavutamiseks, mis alusel, mis isikuandmeid ja mis meetoditega töödeldakse

Uuringu eesmärk on töötada välja teaduspõhine metoodika, mis võimaldab hinnata ja prognoosida elanikkonna haavatavust ning võimalikku abivajadust kahes prioriteetses hädaolukorra stsenaariumis — pikaajalise elektrikatkestuse ja ulatusliku evakuatsiooni korral — piirkondade lõikes. Uuringu eesmärk on toetada hädaolukordade valmisoleku parandamist, riskijuhtimist ja ressursside tõenduspõhist planeerimist ning võimaldada hinnata, millistes piirkondades ja millistel elanikkonnagruppidel võib tekkida suurem abivajadus erineva kiireloomulisusega kriisiolukordades.

Isikuandmete töötlemine toimub teadusuuringu läbiviimise eesmärgil avalikes huvides kooskõlas isikuandmete kaitse üldmääruse (IKÜM) artikli 6 lõike 1 punktiga e ning artikli 9 lõike 2 punktiga j. Uuring viiakse läbi eetikakomitee loa alusel ning andmete töötlemine toimub Statistikaameti teadlaste turvalises töökeskkonnas.

Uuringus kasutatakse olemasolevaid registriandmeid ning uusi isikuandmeid uuringu käigus ei koguta. Andmed pärinevad riiklikest registritest ja infosüsteemidest, sealhulgas Tervisekassa andmekogudest, tervise infosüsteemist (TIS), sotsiaalteenuste ja -toetuste andmeregistrist STAR, sotsiaalkaitse infosüsteemist SKAIS, töövõime hindamise ja töövõimetuse toetuse andmekogust TETRIS, rahvastikuregistrist, hariduse infosüsteemist (EHIS), Politsei- ja Piirivalveameti elamis- ja töölubade registritest, liiklusregistrist, kinnistusraamatust, Eesti Loomaarstide Lemmikloomaregistrist, Töötukassa registritest ning olemasolevatest küsitlusuuringutest (Euroopa Sotsiaaluuring, Päästeameti ja Riigikantselei uuringud).

Töödeldavad isikuandmete kategooriad hõlmavad terviseandmeid (nt kroonilised haigused, hooldusvajadus, puude olemasolu, töövõimega seotud info), sotsiaalkaitse ja toetustega seotud andmeid, haridus- ja töötamise andmeid, demograafilisi andmeid (nt vanus, sugu, elukoha piirkond), leibkonna ja elamistingimustega seotud andmeid, transpordi- ja liikumisvõimekusega seotud andmeid, elamis- ja töölubade andmeid ning hoiakute, teadmiste ja hädaolukordadeks valmisoleku praktikatega seotud andmeid.

Andmete ühendamise viib läbi Statistikaamet oma turvalises keskkonnas. Uurijatele tehakse kättesaadavaks eelnevalt pseudonümiseeritud andmestik ning pseudonüümimise võti ei ole uurijatele ligipääsetav.

Andmeid töödeldakse statistiliste analüüside ja juhendatud masinõppe meetoditega. Uuringus kasutatakse haavatavusindeksite ja komposiitindeksite koostamist, registriandmete sidumist, puuduvate väärtuste imputatsiooni, ruumilist agregeerimist ning juhendatud masinõppemudeleid, sealhulgas otsustuspuid (Decision Trees), juhumetsi (Random Forest), gradientvõimenduse mudeleid (XGBoost, LightGBM) ja närvivõrgupõhist TabNet mudelit. Analüüsi eesmärk on prognoosida piirkondlikke riskitasemeid ja hinnata võimalike abivajajate hulka erinevates hädaolukorra stsenaariumides.

Tulemusi kasutatakse üksnes teaduslikel ja avalikes huvides olevatel eesmärkidel ning avaldatakse ainult agregeeritud kujul piirkondlike hinnangute, kaartide, tabelite ja teaduspublikatsioonidena.

2) Kuidas töötlemiseks kavandatud valikud on eesmärgi saavutamiseks vajalikud ja kohased

Uuringu eesmärk on hinnata elanikkonna haavatavust ja võimalikku abivajadust konkreetsetes hädaolukorra stsenaariumides piirkondade lõikes. Selle eesmärgi saavutamiseks on vajalik kasutada mitmest registrist pärinevaid andmeid, kuna ükski üksik andmekogu ei sisalda piisavat infot haavatavuse eri dimensioonide kohta. Haavatavus kujuneb tervise-, sotsiaalmajanduslike, demograafiliste, elukeskkondlike ja käitumuslike tegurite koosmõjus ning nende tegurite koondumise hindamiseks tuleb andmed erinevatest registritest siduda üheks analüüsitavaks andmestikuks.

Registriandmete kasutamine on eesmärgi saavutamiseks vajalik ja kohane, kuna see võimaldab hinnata ka selliseid elanikkonnarühmi, kelle osakaal küsitlusuuringutes oleks liiga väike usaldusväärsete järelduste tegemiseks. Registriandmed võimaldavad katta kogu elanikkonda ning teha piirkondlikke hinnanguid suurema täpsusega kui ainult küsitlusandmete põhjal. Samuti vähendab olemasolevate registriandmete kasutamine vajadust koguda uusi andmeid otse inimestelt ning vähendab seeläbi andmesubjektide koormust.

Kasutatavad andmekogud on valitud lähtuvalt uurimistöö eesmärgist ja vajadusest mõõta konkreetseid haavatavustegureid. Näiteks tervise- ja hooldusvajadusega seotud haavatavuste hindamiseks kasutatakse tervise- ja sotsiaalkaitse andmeid, sotsiaalmajanduslike riskide hindamiseks töötamise, toetuste ja haridusega seotud andmeid ning liikumis- ja

evakuatsioonivõimekuse hindamiseks elukoha ja transpordiga seotud andmeid. Hoiakute, teadmiste ja valmisoleku praktikate hindamiseks kasutatakse olemasolevaid küsitlusuuringuid agregeeritud kujul. Töödeldavate andmete valik on piiratud ainult nende andmetega, mis on vajalikud uurimisküsimustele vastamiseks ja haavatavuse mudelite loomiseks.

Andmete ühendamine toimub Statistikaameti teadlaste turvalises töökeskkonnas ning ühendamise viib läbi Statistikaamet pseudonümiseeritud kujul. Selline lahendus on vajalik ja kohane, kuna see võimaldab analüüsi läbi viia ilma, et uurijad pääseksid ligi otsestele identifikaatoritele või pseudonüümimise võtmele. Uurijad töötavad üksnes pseudonümiseeritud andmetega ning ligipääs andmetele on piiratud turvalise autentimise ja lepinguliste tingimustega.

Masinõppe mudelite kasutamine on vajalik, kuna uurimistöö eesmärk on prognoosida keerukate ja omavahel seotud haavatavustegurite kuhjumist erinevates hädaolukorra stsenaariumides. Erinevate mudelite (otsustuspuud, random forest, boosting-mudelid, TabNet) katsetamine on põhjendatud, sest nende prognoosivõime ja andmevajadus erinevad ning uuringu üheks eesmärgiks on leida lahendus, mis tagab piisava prognoositäpsuse minimaalse andmekasutusega. Mudelite võrdlemine võimaldab hinnata, milline lähenemine saavutab uurimiseesmärgi kõige väiksema andmemahu ja vähima privaatsusriskiga.

Ruumilise agregeerimise kasutamine (nt $100\text{ m} \times 100\text{ m}$ või $1\text{ km} \times 1\text{ km}$ ruudustik) on vajalik, et tulemused oleksid praktiliselt kasutatavad hädaolukordade planeerimisel ja riskijuhtimisel, säilitades samal ajal andmesubjektide privaatsuse. Üksikisiku tasandi riskiskoore ei avaldata ega visualiseerita ning kõik uuringu tulemused esitatakse üksnes statistiliselt üldistatud kujul. Väikeste rühmade tuvastamise vältimiseks ei avaldata tulemusi kategooriates, kus vaatluste arv jääb alla kehtestatud piirmäära, vaid sellised andmed agregeeritakse suurematesse rühmadesse või jäetakse avaldamata.

3) Ohtude analüüs, sh nende taseme määramine

Käesoleva uuringu puhul tulenevad peamised riskid suuremahuliste registriandmete ühendamisest ning eriliigiliste isikuandmete töötlemisest teadusuuringu eesmärgil. Kuna uuring hõlmab kogu Eesti elanikkonda ning erinevate andmekogude riskasutust, võib isikuandmete töötlemisega kaasneda kõrgendatud risk andmesubjektide õigustele ja vabadustele.

Kaudse taasidentifitseerimise oht — kõrge

Kõige olulisem risk on kaudne taasidentifitseerimine registrite ühendamise tulemusena. Erinevate registrite andmete kombineerimine võib teoreetiliselt võimaldada üksikisikute tuvastamist juhul, kui andmestikus esineb haruldasi tunnuste kombinatsioone või väga väikese suurusega rühmi. Risk suureneb eriti väikeste piirkondade või spetsiifiliste haavatavusprofiilide puhul.

Eriliigiliste isikuandmete töötlemise risk — kõrge

Uuringus kasutatakse terviseandmeid, hooldusvajaduse, puude, töövõime ning sotsiaalkaitsega seotud andmeid, mis kuuluvad eriliigiliste isikuandmete hulka. Selliste andmete töötlemine võib kaasa tuua kõrgendatud riski andmesubjektide privaatsusele ja õigustele, eriti juhul, kui andmeid kasutatakse väljaspool teadusuuringu eesmärki või kui toimuks andmeleke.

Andmeturbe rikkumise või volitamata ligipääsu oht — keskmine

Registriandmete töötlemine ja ühendamine võib kaasa tuua riski, et andmetele pääsevad ligi isikud, kellel puudub selleks õigus, või et andmeid töödeldakse väljaspool lubatud eesmärke. Risk tuleneb eelkõige suurest andmemahust ja mitme andmekogu kasutamisest.

Stigmatiseerimise ja soovimatu sildistamise oht — keskmine

Haavatavuse piirkondlik või rühmapõhine kaardistamine võib kaasa tuua riski, et teatud piirkondi või elanikkonnarühmi tõlgendatakse negatiivses võtmes. Eriti võib see puudutada terviseprobleemide, toetuste saamise või sotsiaalse haavatavusega seotud gruppe.

Automatiseeritud profileerimise väär tõlgendamise oht — madal

Kuigi uuringus kasutatakse masinõppe meetodeid ja riskiskooride prognoosimist, ei tehta üksikisikute suhtes automatiseeritud otsuseid ega looda individuaalseid profile praktiliste sekkumiste tegemiseks. Risk seisneb peamiselt võimaluses, et uuringu tulemusi võidakse ekslikult tõlgendada individuaalse riskihindamisena.

Eesmärgist kõrvalekalduva või liigse andmetöötluse oht — madal kuni keskmine

Kuna uuringus kasutatakse suurt hulka registriandmeid, esineb risk, et töödeldakse rohkem andmeid kui uurimiseesmärgi saavutamiseks vajalik. Risk suureneb juhul, kui andmete valik ei ole piisavalt piiratud konkreetsete haavatavustegurite ja uurimisküsimustega.

4) Milliseid tagatise ja meetmeid ohtude suhtes rakendatakse

Uuringus rakendatakse kogu andmetöötluse vältel tehnilisi, organisatsioonilisi ja metodoloogilisi kaitsemeetmeid, mille eesmärk on vähendada andmesubjektide õigustele ja vabadustele tekkivaid riske ning tagada isikuandmete töötlemise vastavus andmekaitse põhimõtetele. Kuna uuring hõlmab ulatuslikku registriandmete ühendamist ja eriliigiliste isikuandmete töötlemist, pööratakse erilist tähelepanu pseudonümiseerimisele, ligipääsukontrollile, andmete minimaalsusele ning väljundite konfidentsiaalsuse tagamisele.

Pseudonümiseerimine ja identifikaatorite eraldamine

Kõik uuringus kasutatavad registriandmed pseudonümiseerib Statistikaamet enne nende teadlaste töökeskkonda viimist. Pseudonümiseerimise käigus eemaldatakse andmestikust otsesed identifikaatorid ning isikute sidumiseks kasutatav võti säilitatakse eraldi Statistikaametis. Uurijatele ei ole pseudonümimise võti ega otsesed identifitseerivad tunnused kättesaadavad. See vähendab oluliselt üksikisikute tuvastamise riski ka juhul, kui andmestikus kasutatakse mitmest registrist pärinevaid tunnuseid.

Turvaline andmetöötluskeskkond

Andmeid töödeldakse üksnes Statistikaameti teadlaste turvalises töökeskkonnas. Andmetele ligipääs on piiratud ainult uuringuga seotud isikutel ning toimub pärast kasutaja autentimist kaheastmelise turvalahenduse abil (ID-kaart ning kasutajanimi/parool). Ligipääs keskkonnale toimub VPN-ühenduse kaudu. Uuringu jaoks luuakse eraldi töökataloog, millele pääsevad ligi ainult vastava õigusega uurijad. Mikroandmed ei välju üheski uuringu etapis Statistikaameti turvakeskkonnast.

Andmete minimaalsus ja eesmärgipärasus

Uuringus kasutatakse ainult neid andmeid ja tunnuseid, mis on vajalikud uurimiseesmärgi saavutamiseks ning haavatavuse hindamiseks hädaolukordade kontekstis. Registriandmete riskasutamine on piiratud üksnes nende tunnustega, mis on vajalikud haavatavuse ja abivajaduse modelleerimiseks. Mudelite väljatöötamisel võrreldakse erinevaid lähenemisi eesmärgiga leida võimalikult suure prognoosivõimega lahendus minimaalse andmekasutuse juures. Vajadusel kasutatakse tunnuste vähendamist ja andmete agregeerimist, et piirata liigset andmetöötlust.

Agregeeritud väljundid ja taasidentifitseerimise vältimine

Uuringu tulemusi ei avaldata üksikisiku tasandil. Kõik väljundid esitatakse üksnes agregeeritud kujul piirkondlike või rühmapõhiste hinnangutena. Analüüsi käigus hinnatakse erinevaid ruumilise agregeerimise tasemeid (nt 100 m × 100 m või 1 km × 1 km ruudustik), et leida tasakaal analüüsi praktilise kasutatavuse ja privaatsuse kaitse vahel.

Taasidentifitseerimise vältimiseks rakendatakse väikeste rühmade avaldamispiiranguid. Kui mõnes tabelis, piirkonnas või tunnuste kombinatsioonis on vähem kui 20 indiviidi, siis vastavaid tulemusi ei avaldata või agregeeritakse suurematesse kategooriatesse. Samuti välditakse haruldaste tunnuste kombinatsioonide detailsel kujul esitlust kaartidel ja tabelites.

Väljundite kontroll enne väljastamist

Kõik Statistikaameti teadlaste keskkonnast väljastatavad tulemused läbivad enne avaldamist Statistikaameti poolse konfidentsiaalsuskontrolli. Kontrollitakse, et väljundid ei sisaldaks väikese sagedusega kombinatsioone, kaudselt tuvastatavaid andmeid ega muud infot, mis võiks võimaldada üksikisikute identifitseerimist. Teadlaste keskkonnast väljastatakse ainult eelkontrollitud tabelid, joonised ja agregeeritud statistilised tulemused.

Lisaks väljastatakse tulemused Wordi või Exceli dokumentidena, mitte otse statistikapakettide failidena, et vältida olukordi, kus tarkvara väljunditesse võivad jääda peidetud metaandmed või alusandmed.

Lepingulised meetmed

Kõik uurijad sõlmivad Statistikaametiga andmekasutuslepingu, milles sätestatakse andmete kasutamise tingimused, konfidentsiaalsuskohustus ja vastutus andmekaitse nõuete järgimise eest. Andmetele ligipääsu saavad ainult need uurijad, kellel on uuringu läbiviimiseks põhjendatud vajadus.

Uuring viiakse läbi eetikakomitee loa alusel ning andmetöötlus toimub kooskõlas hea teadustava, teaduseetika ning andmekaitse põhimõtetega. Kõik andmetöötluse etapid dokumenteeritakse ning analüüsikood säilitatakse reprodutseeritavuse tagamiseks.

Automatiseeritud otsuste vältimine

Kuigi uuringus kasutatakse masinõppe mudeleid, ei kasutata neid üksikisikute kohta otsuste tegemiseks ega operatiivseks profileerimiseks. Mudelite eesmärk on hinnata piirkondlikke ja rühmapõhiseid haavatavusmustreid teadusuuringu kontekstis. Uuringu tulemused ei too kaasa üksikisikutele õiguslikke tagajärgi ega mõjuta otseselt nende õigusi või kohustusi.

Andmesubjektide koormuse vähendamine

Uuringus ei koguta uusi isikuandmeid otse inimestelt, vaid kasutatakse olemasolevaid registriandmeid ja varasemate küsitlusuuringute andmeid. See vähendab andmesubjektide ajalist ja halduskoormust ning väldib korduvaid andmekogumisi, eriti haavatavate gruppide puhul. Registriandmete kasutamine võimaldab saada vajalikku teaduslikku infot ilma inimesi täiendavalt koormamata.