AEQUITAS

D2.1 Šališkumo ataskaita

Bendrai finansuoja
Europos Sąjunga

AEQUITAS

Turinys

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

AEQUITAS

Įvadas

Viena AEQUITAS projekto užduočių yra sukurti duomenų bazę apie lyties ir rasinį šališkumą medicinoje naudojamuose dirbtinio intelekto (DI) įrankiuose, ypatingą dėmesį skiriant trims ligų grupėms: širdies ir kraujagyslių ligoms, diabetui ir depresijai.

Siekiant įgyvendinti šią užduotį, konsorciumo partneriai pirmiausia surinko įvairius šaltinius apie minėtas šališkumo formas. Kelno universitetinė ligoninė (Universitätsklinikum Köln, UKK), kaip užduoties lyderė ir srities ekspertė, organizavo informacijos rinkimo veiklą ir pateikė struktūrizuotą šabloną, kuriuo partneriai naudojosi žymėdami šaltinius ir užtikrindami, kad reikalinga informacija galėtų būti lengvai perkelta į duomenų bazę.

Šioje ataskaitoje (D2.1) pristatomi teoriniai ir moksliniai pagrindai, kuriais vadovautasi renkant informaciją ir rengiant struktūruotą šabloną, pateikiami atvejų tyrimai, atskleidžiantys skirtingas šališkumo rūšis, taip pat aptariamas poveikis politikos formavimui – kaip biomedicininių DI sistemų sukeliamas šališkumas veikia Europos Sąjungos (ES) pagrindinių teisių chartijos saugomas teises. Taip pat pateikiamas duomenų rinkimo veiklos aprašymas, susiejimo šablonas, AEQUITAS partnerių surinktų šaltinių sąrašas ir kita papildoma medžiaga. Likusi ataskaitos dalis struktūruota taip:

Pirmiausia pateikiamas teorinis pagrindas – įvadas į medicinoje naudojamus DI įrankius ir šališkumo sampratą kompiuterinėse sistemose bei medicinoje. Pirmiausia nagrinėjama medicina: pirma, aptariama, kaip rasinis ir lyčių šališkumas pasireiškia sveikatos priežiūroje, antra – kaip bioetika sprendžia etinius klausimus, kylančius medicinos praktikoje ir biomedicininiuose tyrimuose, trumpai pristatant keturis bioetikos principus (autonomiją, nekenkimą, gerovės siekimą ir teisingumą).

Toliau pereinama prie DI srities ir pristatomos skirtingos šališkumo rūšys, pasireiškiančios DI sistemose ir mašininio mokymosi / dirbtinio intelekto (MM/DI) vystymo grandinėje. Kiekviena šališkumo rūšis iliustruojama pavyzdžiais ir atvejų tyrimais apie lyčių ir rasinį šališkumą bei jų poveikį visuomenei, susijusį su trimis

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

AEQUITAS projekto tikslinėmis ligomis (širdies ir kraujagyslių ligomis, diabetu ir depresija). Šie pavyzdžiai parinkti iš partnerių surinktų šaltinių po T2.2 užduoties įgyvendinimo. Tais atvejais, kai surinkta medžiaga aiškiai neatskleidė konkretaus DI šališkumo tipo, pateiktas analogiškas pavyzdys iš kitos medicinos srities, kurį galima apibendrinti AEQUITAS tikslinėms ligoms. Atvejų analizės papildytos papildomais moksliniais šaltiniais.

Paskutiniame – politikos pasekmių – skyriuje, aptariama, kaip skirtingos DI šališkumo rūšys veikia pagrindines teises, saugomas ES chartijoje – ypač žmogaus orumą, lygybę prieš įstatymą, nediskriminavimą, teisę į asmens neliečiamumą, teisę į sveikatos priežiūrą, duomenų apsaugą ir į veiksmingą teisinę gynybą – bei kokios apsaugos priemonės gali būti taikomos atitikties vertinimo, galutinės stebėsenos ir viešųjų pirkimų procesuose.

Ataskaita baigiama literatūros sąrašu ir šiais priedais:

1 priedas: Šaltinių rinkimo ir susiejimo metodas – pateikiamas susiejimo šablonas ir aprašomas rinkimo bei vertinimo procesas T2.1 ir T2.2 užduotyse.

2 priedas: Papildoma medžiaga T2.1 ir T2.2 užduotims – partnerių susitikimų skaidrės, parengtos UKK.

3 priedas: AEQUITAS partnerių surinktų šaltinių sąrašas.

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

DI medicininės taikomosios sistemos

Pastaraisiais metais sparčiai augantis DI naudojimas padarė didelę įtaką medicinai, įskaitant skaitmenizuotą duomenų rinkimą, mašininį mokymąsi ir skaičiavimo infrastruktūrą (Yu et al., 2018). Ypač giliojo mokymosi algoritmų įdiegimas, tokiose srityse kaip kompiuterinė rega ir natūralios kalbos apdorojimas, iš esmės pakeitė kompiuterinių sistemų taikymą radiologijoje, patologijoje, kardiologijoje, diabetologijoje, psichiatrijoje, onkologijoje ir kitose srityse. (Esteva et al., 2019; Koteluk et al., 2021; Rajpurkar et al., 2022; Gou et al., 2024). Pasaulio sveikatos organizacija (PSO) išskiria šias DI sistemų taikymo sritis sveikatos priežiūroje: diagnostika ir prognozavimu pagrįsta diagnostika, klinikinė priežiūra, moksliniai tyrimai ir vaistų kūrimas, sveikatos sistemų valdymas ir planavimas, visuomenės sveikata ir jos stebėsena, sveikatos stiprinimas, ligų prevencija, prognozavimo pagrindu vykdoma stebėsena, pasirengimas ekstremalioms situacijoms ir reagavimas į protrūkius (Pasaulio sveikatos organizacija, 2021).

Tačiau DI diegimas medicinoje kelia ir iššūkių: įgyvendinimo sunkumus, įskaitant pasitikėjimą modeliais ir duomenų ribotumą, atskaitomybės problemas, įtraukiant reglamentavimo ir atsakomybės priskyrimo klausimus, bei sąžiningumo užtikrinimą – etišką duomenų naudojimą, teisingą naudos paskirstymą, šališkumo nustatymą ir mažinimą (Rajpurkar et al., 2022).

AEQUITAS projektas nagrinėja lyčių ir rasinio šališkumo atvejus širdies ir kraujagyslių ligų, diabeto ir depresijos srityse. DI medicininės sistemos padeda širdies ir kraujagyslių ligų priežiūrai per klinikinių sprendimų palaikymą, nuotolinę mediciną, rizikos vertinimą, individualizuotą terapiją, prognozinę analizę ir nuotolinę stebėseną (Bernstein et al., 2025; Naskar et al., 2025), taip pat gerina diabeto kontrolę (įskaitant pacientų stebėseną ir savikontrolę), diagnostiką, gydymą ir prevenciją (Contreras & Vehi, 2018; Khalifa & Albadawy, 2024; Naskar et al., 2025; Sheng et al., 2024). Depresijos srityje DI naudojamas patikrai, diagnostikai ir gydymui (Alhuwaydi, 2024) ypač nustatymui ir patikrai pasitelkiant didelius kalbos modelius (Cao et al., 2025;

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

Kumari et al., 2025; Mao et al., 2023; Wang et al., 2025). Visose minėtose srityse yra šališkumo iššūkių, pavyzdžiui, dėl širdies ir kraujagyslių ligų, diabeto ir depresijos, žr. atitinkamai (van Assen et al., 2024), (Cronjé et al., 2023), (Dang et al., 2024).

Tokie iššūkiai kaip šališkumas, tiek medicinoje, tiek dirbtiniame intelekte, sprendžiami derinant bioetiką ir dirbtinio intelekto etiką. Kitame skyriuje pateikiame trumpą šių dviejų taikomosios etikos sričių, kurios buvo teorinis ir mokslinis pagrindas kuriant struktūruotą šabloną.

Etika ir šališkumas medicinoje ir DI

Bioetika ir šališkumas medicinoje

Šališkumas medicinoje yra plačiai dokumentuotas. Pavyzdžiui, (Hammond et al. 2021) kognityvinis šališkumas – tai sisteminės mąstymo klaidos, atsirandančios dėl žmogaus informacijos apdorojimo ribotumų ar netinkamų mąstymo modelių (FitzGerald and Hurst 2017) numanomas šališkumas – tai nesąmoningos asociacijos, lemiančios neigiamą asmens vertinimą pagal nereikšmingas savybes, tokias kaip rasė ar lytis.

Rasinis šališkumas medicinoje išsamiai nagrinėtas JAV kontekste, pavyzdžiui nustatyta, kad afroamerikiečiai ir kitų mažumų atstovai gauna mažiau medicininių procedūrų ir žemesnės kokybės priežiūrą: jiems rečiau taikomas intensyvus gydymas, rečiau atliekamos operacijos ir rečiau skiriamos specialistų konsultacijos nei baltaodžiams pacientams (Bowser, 2001; Williams & Wyatt, 2015).

Lyčių šališkumas siejamas su „lyčių aklumu“ ir stereotipinėmis nuostatomis apie vyrus ir moteris (Hamberg, 2008), taip pat su bendru žinių trūkumu apie moters organizmo veikimą ir biologinius skirtumus nuo vyro organizmo. Pavyzdžiui, sunkios būklės 50 metų ir vyresnės moterys rečiau nei vyrai patenka į intensyvios terapijos skyrius (Bierman, 2007) ir netgi laboratorinių pelių patinų eksperimentiniai modeliai yra naudojami dažniau nei patelių eksperimentiniai modeliai klinikiniuose ir chirurginiuose biomedicininiuose tyrimuose (Yoon et al., 2014).

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

AEQUITAS

Svarbu pažymėti, kad LGBT+ asmenys, taip pat patiria diskriminaciją sveikatos priežiūros prieinamumo srityje ir stereotipizavimą, kuris nepaliečia heteroseksualių asmenų. Šie socialiniai ir kultūriniai veiksniai įtvirtina diskriminaciją ir daro įtaką sveikatai. Pavyzdžiui, JAV atliktas tyrimas, pagrįstas 2013–2014 m. Nacionalinės sveikatos apklausos (NHIS) duomenimis, parodė, kad LGBT+ suaugusieji, palyginti su heteroseksualiais asmenimis, nurodė didesnį prastos sveikatos, funkcinių apribojimų, didelio psichologinio streso ir sunkumų gauti sveikatos priežiūros paslaugas lygį. Šią nelygybę lemia mažumų grupių patiriamas stresas ir daugialypė visuomenės marginalizacija (Liu et al., 2023).

Tuo pat metu medicina, kaip disciplina, nuo seniausių laikų grindžiama aukštais etikos standartais (Baker & McCullough, 2008). Daugelį amžių visuomenėje egzistuoja lūkestis, kad gydytojas laikysis profesinės atsakomybės etikos taisyklių, nustatytų jo profesijos standartų. Tai atsispindi profesinėse normose – nuo Hipokrato priesaikos, siekiančios apie 400 metus prieš mūsų erą (Miles, 2005), iki Ženevos ir Helsinkio deklaracijų (Tröhler, 2008). Kaip pažymi (Vevaina et al., 1993), gydytojai, dėl visuomenės investicijų į jų išsilavinimą (tiek finansinių, tiek per visuomenės narių dalyvavimą kaip mokymo medžiagos gydytojų rengimo ir profesinės veiklos metu) bei dėl faktiškai suteikiamos profesinės monopolijos, įtvirtinamos licencijavimo sistema, yra įpareigoti laikytis etikos kodo.

Biomedicinos etika (arba bioetika) yra praktinės (taikomosios) etikos sritis, nagrinėjanti etinius klausimus, kylančius medicinos praktikoje ir biomedicininiuose tyrimuose (Vevaina et al., 1993). Biomedicinos etika remiasi keturiais principais, kuriuos apibrėžė Beauchamp ir Childress (Beauchamp & Childress, 2019):

1. Autonomija: pagarba autonomiškų asmenų gebėjimui priimti sprendimus. Autonomijai būtinos dvi bendrosios sąlygos: laisvė, pasireiškianti nepriklausomumu nuo kontroliuojančių įtakų, ir veikimo geba – tai yra gebėjimas sąmoningai ir tikslingai veikti.

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

AEQUITAS

2. Nekenkimas: pareiga vengti žalos sukėlimo.
3. Gerovės užtikrinimas: pareiga imtis aktyvių veiksmų padėti kitiems – konkrečiai, užkirsti kelią žalai ar blogiui, juos šalinti ir skatinti gerovę.
4. Teisingumas: naudos, rizikų ir sąnaudų paskirstymas teisingai. Teisingumas suprantamas kaip sąžiningas, lygiavertis ir tinkamas požiūris į asmenis ir grupes, atsižvelgiant į sveikatos priežiūros ir mokslinių tyrimų netolygumus, susijusius su rase, etnine kilme, lytimi ir socialine padėtimi.

DI etika ir šališkumas

Dirbtinio intelekto įdiegimas ir spartus jo taikomųjų sprendimų vystymasis iškėlė įvairių etinių klausimų (Christoforaki & Beyan, 2022), tarp kurių ypač svarbūs yra diskriminacijos ir šališkumo klausimai.

Todėl susiformavo DI etika kaip praktinės (taikomosios) etikos sritis, apimanti „vertybių, principų ir metodų visumą, kuri, remdamasi plačiai pripažintais gėrio ir blogio standartais, skatina etišką elgesį kuriant ir naudojant DI technologijas“ (Leslie, 2019, p. 3).

DI etika remiasi tiek bioetika (keturiais pagrindiniais principais, pristatytais aukščiau), tiek žmogaus teisių diskursu. Pastarasis apima teisę į lygią laisvę ir orumą pagal įstatymą, pilietinių, politinių ir socialinių teisių apsaugą, visuotinį asmens teisinio subjektiškumo pripažinimą bei teisę laisvai ir nevaržomai dalyvauti bendruomenės gyvenime (Leslie, 2019).

Keturi bioetikos principai, papildyti paaiškinamumo principu, DI kontekste (Floridi et al., 2018) apibrėžiami taip:

Autonomija – žmogaus galia nuspręsti ar priimti sprendimą, ar perduoti ji sistemai, kartu suvokiant per didelio delegavimo mašinoms riziką.

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

AEQUITAS

Nekenkimas – žalos prevencija, tiek kylančios iš žmogaus ketinimų, tiek iš nenumatyto sistemų elgesio.

Gerovės užtikrinimas – gerovės skatinimas, orumo išsaugojimas ir tvarumo palaikymas.

Teisingumas – esamos nesąžiningos diskriminacijos prevencija ir šalinimas, naujos žalos vengimas bei vienodas DI teikiamos naudos paskirstymas.

Paaiškinamumas – DI sprendimų priėmimo procesų suprantamumas ir atskaitomybė.

Todėl DI etika taip pat susiformavo kaip principų sistema, grindžiama keturiais klasikiniais medicinos etikos principais bei kitais požiūriais, kuriuos apibendrino (Christoforaki & Beyan, 2022). Tačiau, pažymima (Mittelstadt, 2019), kad lyginant su medicina, DI vyątyma srityje trūksta: (1) bendrų tikslų ir fiduciarinių pareigų (2) nusistovėjusios profesinės istorijos ir normų, (3) patikrintų metodų, leidžiančių principus paversti praktika, ir (4) tvirtų teisinių bei profesinės atskaitomybės mechanizmų. Visa tai silpnina principinio požiūrio veiksmingumą. Be to, Europos Sąjungoje DI kūrimą ir naudojimą reglamentuoja sudėtinga teisės aktų sistema, įskaitant antidiskriminacinius įstatymus, tačiau ši tema nepatenka į šios ataskaitos apimtį.

Kalbant apie žmogaus teises, pagal 2018 m. Europos Tarybos finansuotą ataskaitą, (Committee of experts on internet intermediaries (MSI-NET), 2018), žmogaus teisės, kurias ypač veikia algoritmai ir automatizuoto duomenų apdorojimo metodai, apima:

- teisę į teisingą bylos nagrinėjimą ir tinkamą teisinį procesą,
- privatumo ir duomenų apsaugos teisę,
- saviraiškos laisvę,
- teisę į veiksmingą teisinę gynybą,
- susirinkimų ir asociacijų laisvę,
- diskriminacijos draudimą,

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

AEQUITAS

- socialines teises ir prieigą prie viešųjų paslaugų,
- teisę į laisvus rinkimus.

Šališki algoritmai aiškiai įvardijami kaip galimi diskriminuojantys veiksniai visuomenės grupių atžvilgiu dėl jų amžiaus, seksualinės orientacijos, rasės, lyties ar socio-ekonominės padėties (Committee of experts on internet intermediaries (MSI-NET), 2018, p. 27). Be to, Europos Tarybos pagrindų konvencijoje dėl dirbtinio intelekto ir žmogaus teisių, demokratijos ir teisinės valstybės principais nustatyta, kad valstybės narės „turi priimti arba išlaikyti priemones, užtikrinančias, kad DI sistemų gyvavimo ciklo veiklos gerbtų lygybę, įskaitant lyčių lygybę, ir diskriminacijos draudimą pagal taikomą tarptautinę ir nacionalinę teisę“ (Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, 2024, p. 4).

Pilietinės visuomenės organizacijos (PVO), kaip sveikatos priežiūros ekosistemos suinteresuotosios šalys (Vayena et al., 2018), gali atlikti reikšmingą vaidmenį nustatant ir mažinant DI šališkumą bei formuojant DI valdymą – skatindamos etišką DI kūrimą, užtikrindamos suinteresuotųjų šalių atskaitomybę, šviesdamos visuomenę, atstovaudamos marginalizuotoms bendruomenėms, dalyvaudamos politikos ir reguliavimo sistemų formavime bei stiprindamos bendradarbiavimą tarp valdžios institucijų, technologijų įmonių ir visuomenės (Korir, 2024).

Remiantis šiuo teoriniu pagrindu, kuriami įvairūs techniniai sprendimai šališkumui mažinti. Kitame skyriuje pateikiama DI sukeltų šališkumų klasifikacija, kuri buvo naudojama kaip pagrindas struktūruotam šablonui sudaryti, daugiausia dėmesio skiriant jų poveikiui lyčių ir rasinei diskriminacijai. Žmogaus mąstymo šališkumai (Hofmann, 2023), pavyzdžiui, kognityviniai šališkumai, tokie kaip patvirtinimo ar prieinamumo šališkumas, nors ir labai svarbūs medicinoje, šiame projekte nepriskiriami nagrinėjamai apimčiai.

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

Šališkumas DI sistemose

Šališkumas kompiuterinėse sistemose apibrėžiamas (Friedman & Nissenbaum, 1996, p. 332) kaip reiškinys, kai kompiuterinės sistemos sistemingai ir nesąžiningai diskriminuoja tam tikrus asmenis ar asmenų grupes kitų naudai. Sistema diskriminuoja nesąžiningai, jei ji atima galimybę ar naudą arba priskiria nepageidaujamą rezultatą asmeniui ar grupei dėl nepagrįstų ar netinkamų kriterijų.

Pagal (Friedman & Nissenbaum, 1996), kompiuterinę klasifikaciją skiriamos trys šališkumo sistemose kategorijos: iš anksto egzistuojantis šališkumas, techninis šališkumas ir atsirandantis šališkumas. Toliau kiekviena šališkumo rūšis aptariama atskirai ir iliustruojama mokslinėje literatūroje pateiktais atvejų tyrimais.

Išankstinis šališkumas

Iš anksto egzistuojantis šališkumas kyla iš socialinių institucijų, praktikų ir nuostatų, kurios jau egzistuoja nepriklausomai nuo kuriamos sistemos. Šis šališkumas į sistemą gali būti įtrauktas sąmoningai arba nesąmoningai, net ir tada, kai sistemos vystytojai stengiasi jo išvengti.

Atvejo analizė: širdies ir kraujagyslių ligų diagnostika moterims

Širdies ir kraujagyslių ligos ilgą laiką buvo laikomos „vyrų ligomis“, ir toks požiūris prisidėjo prie moterų nepakankamos diagnostikos ir gydymo. Sisteminės apžvalgos (Al Hamid et al., 2024) rodo, kad moterims šios ligos diagnozuojamos rečiau, nes jų simptomai dažnai būna silpnesni nei vyrų arba klaidingai priskiriami virškinimo ar nerimo sutrikimams. Dėl to moterims rečiau skiriami diagnostiniai tyrimai ir vaistai, jos rečiau nukreipiamos pas kardiologus ir yra rečiau hospitalizuojamos. Net ir patekusios į ligoninę, jos rečiau sulaukia invazinių koronarinės kraujotakos intervencijų. Taip pat nustatyta, kad gydytojai – ypač vyrai – linkę nepakankamai įvertinti moterų rizikos veiksnius. Atsižvelgiant į tai, kad moterys vis dar yra nepakankamai atstovaujamos

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

AEQUITAS

kardiologijos srityje (Fatunde et al., 2025), galima daryti išvadą, jog esami šališkumai lemia prastesnę jų sveikatos priežiūrą.

Kadangi DI sistemos mokomos naudojant realios praktikos duomenis, DI pagrindu veikianti širdies ir kraujagyslių ligų diagnostikos sistema įtraukia šiuos šališkumus ir gali sistemingai diskriminuoti moteris, nepriklausomai nuo techninių įgyvendinimo sprendimų.

Techninis šališkumas

Techninis šališkumas atsiranda dėl techninių apribojimų ar projektavimo sprendimų, kai žmogiškos sąvokos pritaikomos kompiuteriniam apdorojimui – kiekybiškai įvertinant kokybinius reiškinius, diskredituojant tęstinius procesus ar formalizuojant neformalius reiškinius. Be to, algoritmų atskyrimas nuo jų veikimo konteksto gali lemti nevienodą skirtingų grupių vertinimą skirtingomis sąlygomis.

Atvejo analizė: insulto rizikos prognozavimo modelių tikslumas juodaodžių ir baltaodžių populiacijose

(Hong et al., 2023) retrospektyvinis tyrimas, kuriame buvo lyginamas insulto rizikos prognozavimo modelių tikslumas, parodė, kad visi nagrinėti algoritmai prasčiau išskyrė riziką juodaodžių pacientų grupėje nei baltaodžių grupėje. Situacija, tyrėjų teigimu, gali būti susijusi su tuo, kad duomenyse nebuvo užfiksuoti svarbūs rizikos veiksniai – draudimo tipas, kalbos barjerai ir kiti veiksniai, susiję su nevienoda prieiga prie sveikatos priežiūros paslaugų. Kitaip tariant, duomenys buvo atskirti nuo jų socialinio ir ekonominio konteksto. Tuo pačiu visi aukščiau išvardyti rizikos veiksniai yra sudėtingos konstrukcijos, kurias sunku atvaizduoti kompiuteriams tinkama forma. Be to, galima pridurti, kad pažangiausi šiuolaikiniai DI algoritmai iš prigimties naudoja neskaidrius ir neaiškius kriterijus dideliam tikslumui pasiekti (Knight, 2017). Dėl to dažnai net jų

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

AEQUITAS

vystytojai negali paaiškinti, kaip tiksliai veikia algoritmai, ir todėl negali užtikrintai kontroliuoti ar minėti socialiniai ir ekonominiai veiksniai iš tiesų yra įtraukiami į vidinį DI sistemos sprendimų priėmimo procesą.

Iškylantis šališkumas

Iškylantis šališkumas pasireiškia baigus sistemos vystymą ir specialistams ją naudojant. Jis atsiranda dėl kintančių visuomenės žinių, kurios negali būti arba nėra įtraukiamos sistemos vystymo metu, arba dėl vartotojų, kurių žinios ar kultūrinės vertybės skiriasi nuo tų, kurios buvo numanomos kuriant sistemą.

Atvejo analizė: duomenų rinkinių poslinkiai

Duomenų rinkinio poslinkis – tai neatitikimas tarp mokymo ir testavimo duomenų pasiskirstymo algoritmo kūrimo metu, kuris gali lemti nevienodą veikimą skirtinguose pogrupiuose (Chen et al., 2023).

Pavyzdžiui, odos vėžio nustatymo srityje daugelis naudojamų vaizdų rinkinių yra iš šalių, kuriose populiacija pagrinde yra šviesiaodė (Guo et al., 2021), todėl tam tikros demografinės grupės yra nepakankamai atstovaujamos. Tokiais duomenimis apmokyti DI modeliai prasčiau veikia juos taikant įvairesnės sudėties populiacijose ir diskriminuoja tamsiaodžius asmenis. Duomenų rinkinių sudarymas ir žymėjimas yra sudėtingas ir brangus procesas, todėl kuriant DI sistemas žemų ir vidutinių pajamų šalyse dažnai naudojamasi viešaisiais duomenų rinkiniais, kurie neatspindi jų populiacijos struktūros. Tai sukuria neatitikimą tarp šaltinio ir tikslinės populiacijos. Panašūs neatitikimai gali atsirasti ir aukštų pajamų šalyse dėl migracijos ar rasės saviidentifikacijos pokyčių. Be to (Chen et al., 2023), vis plačiau pripažįstama, kad rasė yra socialinis konstruktas ir genetinis kintamumas grupių viduje yra didesnis nei tarp jų, todėl senos klasifikacijos nepakankamai tiksliai apibūdina realias žmonių grupes ir gali užmaskuoti kultūrinius, istorinius bei socialinius veiksnius, svarbius teisingumo vertinimui.

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

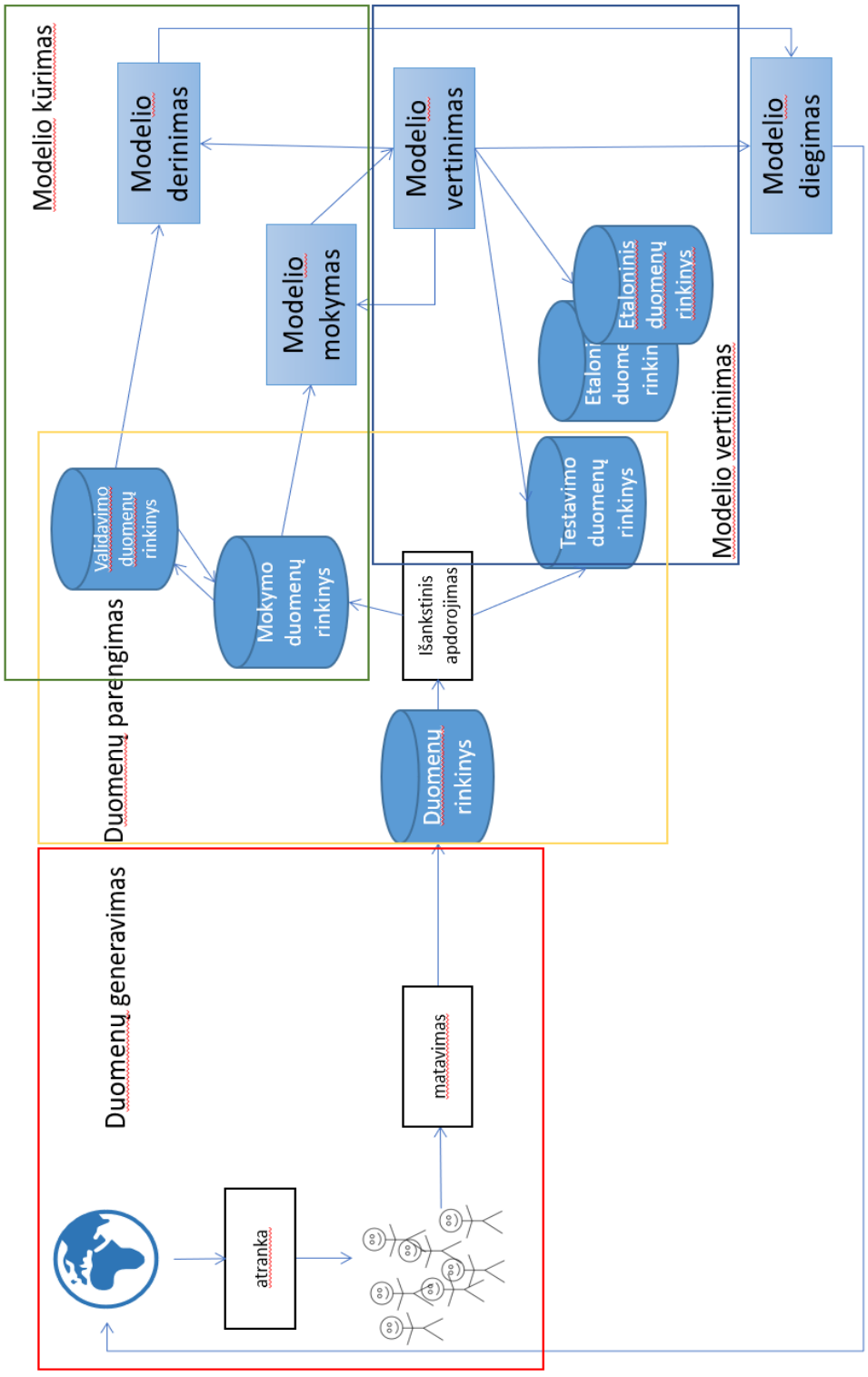
Šališkumo rūšys, būdingos MM/DI vystymo grandinei

Nors tai, kas išdėstyta aukščiau, galioja visoms kompiuterinėms sistemoms, DI taikomosioms sistemoms būdingi specifiškesni reikalavimai, todėl mums buvo reikalinga detalesnė klasifikacija. Todėl nusprendėme remtis (Suresh & Guttag, 2020), pateikta šališkumo klasifikacija, nes ji identifikuoja šališkumo tipus kiekviename MM/DI vystymo grandinės etape, kaip pavaizduota 1 paveiksle.

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

AEQUITAS



1 pav. MM/DI veikimo (vystymo) etapas. Paveikslas adaptuotas pagal Suresh & Guttag (2020).

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

AEQUITAS

Įprasta MM/DI vystymo grandinė gali būti apibūdinama taip:

Duomenų generavimas. MM/DI sistemos vystymas prasideda nuo duomenų generavimo. Tai apima duomenų rinkimą ir parengimą, siekiant sudaryti duomenų rinkinį sistemai mokyti. Reikia nustatyti tikslinę populiaciją ir iš jos surinkti imtį. Tuomet apibrėžiami ir išmatuojami taikymui svarbūs požymiai ir (arba) duomenys sužymimi tinkamomis etiketėmis. Tai brangus ir ilgas procesas, todėl praktikoje dažnai naudojami jau egzistuojantys (vieši arba įsigyti) duomenų rinkiniai.

Duomenų parengimas. Šiame etape duomenų rinkinys padalijamas į tris dalis: mokymo duomenis, naudojamus modeliui mokyt; validavimo duomenis, naudojamus modelio tinkamumui vertinti derinant modelio hiperparametrus (modelio parametrai, kurių negalima išmokti iš duomenų, pvz., sluoksnių ir neuronų skaičius neuroninio tinklo modelyje). Šiame etape gali tekti iš anksto apdoroti duomenis (pvz., išvalyti, normalizuoti); ir sukurti bandymų duomenų rinkinį – duomenų dalį, naudojamą galutiniam modeliui įvertinti, kuris tampa auksiniu standartu, kai modelis yra visiškai apmokytas.

Modelio vystymas. Šiame etape modelis mokomas naudojant mokymo duomenis ir tobulinamas taikant hiperparametrus validavimo duomenims.

Modelio vertinimas. Išmokytas modelis vertinamas naudojant testavimo duomenis, o kartais ir nepriklausomai sudarytus standartinius duomenų rinkinius, kurie naudojami modelio patikimumui įrodyti arba palyginimui su kitais metodais.

Modelio diegimas. Modelis pritaikomas realioje aplinkoje. Šio testavimo rezultatai gali lemti pakeitimus ir iškelti taisytinus aspektus ankstesniuose vystymo etapuose.

Atsižvelgdami į aukščiau aprašytus MM/DI proceso etapus, taikome (Suresh & Guttag, 2020) šališkumo klasifikaciją. Konkrečiai, jie išskiria šias šališkumo kategorijas: istorinis, reprezentacijos šališkumas, matavimo, agregavimo, mokymosi, vertinimo ir diegimo šališkumas. Tolesniuose poskyriuose apibrėžiame aukščiau išvardytus šališkumus ir pateikiame atvejų analizes iš surinktų šaltinių.

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

Istorinis šališkumas

Istorinis šališkumas atitinka išankstinį šališkumą, kaip apibrėžta (Friedman & Nissenbaum, 1996), kuris apima jau duomenyse egzistuojančius išankstinius nusistatymus ir stereotipus. (Calderone, 1990) aprašo pavyzdį, kuriame nagrinėjama, ar skausmą malšinančių ir raminamųjų vaistų vartojimo dažnis pacientams po operacinio vainikinių arterijų šuntavimo, skiriasi priklausomai nuo jo lyties ir amžiaus. Rezultatai parodė, kad vyrams bei 61-erių metų ir jaunesniems pacientams skausmą malšinantys vaistai buvo skiriami žymiai dažniau nei moterims bei 62-ejų metų ir vyresniems pacientams, kuriems raminamieji vaistai buvo skiriami žymiai dažniau. Atvejo analizė apie insulto rizikos prognozavimo modelių prognozių tikslumą juodaodžių ir baltaodžių populiacijose tai parodo poskyryje „Iš anksto egzistuojantis šališkumas“; tačiau mes pristatysime kitą atvejo tyrimą, kuriame demonstruojamas istorinis šališkumas, susijęs su dirbtinio intelekto naudojimu psichikos sveikatos srityje.

Atvejo analizė: dirbtinis intelektas psichikos sveikatoje ir kalbos modelių šališkumas

(Straw & Callison-Burch, 2020) pateikia sisteminę literatūros apžvalgą apie neurolingivstinį programavimą (NLP) naudojimą psichikos sveikatos srityje, siekiant nustatyti, kaip šie šališkumai gali padidinti sveikatos nelygybę. Dirbtinio intelekto modeliai, kurie naudoja NLP psichikos sveikatos profiliui nustatyti, renka didelius išraiškingos kalbos duomenų rinkinius, paprastai gaunamus iš socialinės žiniasklaidos, internetinių forumų, tinklaraščių ir pokalbių kambarių. Tačiau šiems duomenims įtakos turi ir žmogaus asmeninė patirtis bei socialinis kontekstas.

Kalbant konkrečiai apie lytį ir kalbą, yra išsami bibliografija (anglų kalba), apibendrinta (Pennebaker et al., 2003), kurioje atskleidžiami moterų ir vyrų žodžių vartojimo skirtumai. Pavyzdžiui, moterys vartoja mažiau kategorišką kalbą, pasireiškiančią didesniu mandagumu, mažiau keiksmažodžių, daugiau stiprinančių žodžių (pvz., „tikrai“, „taip“) ir daugiau ribojančių žodžių (t. y. paaiškinimų ar neužtikrintų žodžių, tokių kaip „kaip ir“, „galbūt“ arba „gal“). Kita vertus, vyrai buvo apibūdinami kaip aiškūs, tikslūs ir mažiau emocingi savo kalboje, kuriai būdinga kiekybė, vertinamieji

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

AEQUITAS

būdvardžiai (pvz., „geras“, „kvailas“), elipsiniai sakiniai („Puikus paveikslėlis.“) ir nuorodos į „aš“. Kaip pažymi autoriai, šie skirtumai atitinka sociologinę lyčių skirtumų sistemą, tačiau juos taip pat galima paaiškinti alternatyviomis prielaidomis, pavyzdžiui, didesniu moterų socialiniu įsitraukimu.

Kalbant apie psichinę sveikatą, vyrai ir moterys skirtingai rašo atsisveikinimo laiškus, kuriuose išreiškia suicidinių minčių sukeltą nerimą; moterys internalizuoja neigiamas emocijas, o vyrai reiškia didėjantį pyktį (Straw & Callison-Burch, 2020). Dirbtinio intelekto sistema, skirta vienos lyties asmenų psichikos sveikatos sutrikimams nustatyti, gali būti netinkama kitai lyčiai (ir tai dar vertinant lytį binariniame kontekste, kuris iš esmės eliminuoja didelę visuomenės dalį).

Reprezentatyvumo šališkumas

Reprezentatyvumo šališkumas atsiranda tada, kai vystymo imtis duomenų rinkimo etape nepakankamai atspindi tam tikrą populiacijos dalį. Tai gali kilti šiais atvejais: apibrėžiant tikslinę populiaciją, jei ji neatitinka realios naudojimo populiacijos; apibrėžiant tikslinę populiaciją, jei joje yra nepakankamai atstovaujamų grupių; imant imtį iš tikslinės populiacijos, jei atrankos metodas yra ribotas arba netolygus. Reprezentatyvumo šališkumas lemia prastą rezultatų apibendrinimą tam tikros naudotojų populiacijos dalies atžvilgiu.

Tipinis reprezentatyvumo šališkumo pavyzdys susijęs su odos vėžio nustatymu, nes daugelyje vaizdų duomenų rinkinių tam tikros demografinės grupės yra nepakankamai atstovaujamos, todėl mašininio mokymosi modeliai yra mokomi daugiausia naudojant šviesiaodžių asmenų atvaizdus (Guo et al., 2021). Atsižvelgiant į AEQUITAS projekto tikslines ligas, pateikiame reprezentatyvumo šališkumo rasės atžvilgiu 2 tipo diabeto kontekste atvejo analizę.

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

Atvejo analizė: rasinio šališkumo vertinimas 2 tipo diabeto rizikos prognozavimo algoritmuose

Remiantis (Cronjé et al., 2023), remiantis JAV populiacijos duomenimis, nepaisant santykinai mažesnės rizikos, ne ispanų kilmės baltieji asmenys išlieka perteklinai atstovaujami diabeto rizikos prognozavimo literatūroje. Kitoje apžvalgoje apie etninį ir rasinį teisingumą taikant DI diabeto valdymui nustatyta, kad tarp straipsnių, kuriuose buvo pateikti duomenys apie rasę, vidutinis pasiskirstymas buvo 69,5 % baltaodžių, 17,1 % juodaodžių ir 3,7 % azijiečių, ir tik dviejuose straipsniuose buvo nurodytas JAV vietinių tautų atstovų dalyvavimas (Pham et al., 2021).

Yra gerai dokumentuota, kad diabeto pasekmių netolygumus daugiausia lemia sudėtingi, tarpusavyje susiję socialiniai sveikatos veiksniai, įskaitant prieigą prie sveiko maisto, sveikatos priežiūros kokybę, sveikatos draudimo statusą, švietimo barjerus ir skirtingus technologijų įsisavinimo lygius. Šios pasekmės apima didesnį komplikacijų dažnį ir prastesnę glikemijos kontrolę tarp marginalizuotų ir mažas pajamas gaunančių populiacijų (Alipour & Alipour, 2025).

Dėl to DI sistema, apmokyta su esamais duomenų rinkiniais, prastai apibendrintų rezultatus, o tai lemtų šališkus prognozavimo modelius, kurie, pavyzdžiui, prevencinių veiksmų srityje galėtų pirmenybę teikti tam tikrų rasinių grupių asmenis.

Matavimo šališkumas

Matavimo šališkumas atsiranda tada, kai pasirenkami, renkami arba apskaičiuojami požymiai ir žymos, naudojami prognozavimo uždavinyje, ypač kai naudojamas pakaitinis rodiklis (tai yra konstrukto, kuris nėra tiesiogiai užkoduotas ar stebimas, apytikris atitikmuo). Pavyzdys pateikiamas (Obermeyer et al. 2019), tyrime, kuriame sveikatos priežiūros išlaidos buvo naudojamos kaip pakaitinis rodiklis siekiant prognozuoti ir reitinguoti, kuriems pacientams papildoma priežiūra būtų naudingiausia, ir tai lėmė rasinę diskriminaciją. Sveikatos priežiūros išlaidos yra netinkamas sveikatos

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

AEQUITAS

poreikio pakaitinis rodiklis, nes juodaodžiai pacientai, susidurdami su neproporcingai dideliu skurdo lygiu, dažnai sveikatos priežiūrai išleidžia mažiau nei baltaodžiai. Dėl šio šališkumo algoritmas klaidingai padarė išvadą, kad juodaodžiai pacientai yra sveikesni nei tokios pat būklės baltaodžiai pacientai, todėl, siekiant gauti sveikatos priežiūros paslaugas, jiems priskyrė žemesnio prioriteto grupę.

Kiti matavimo šališkumo šaltiniai gali atsirasti tada, kai matavimo metodas skiriasi tarp grupių, pavyzdžiui, kai dvi grupės stebimos dėl to paties elgesio, tačiau viena iš jų stebima griežčiau arba dažniau nei kita. Panašiai matavimo tikslumas gali skirtis tarp grupių, o tai medicininėse taikomosiose srityse gali lemti sistemiškai didesnius klaidingos diagnozės arba nepakankamos diagnostikos rodiklius tam tikrose grupėse. Pavyzdžiui, gydytojai dažniau nuvertina juodaodžių pacientų patiriamą skausmą, lyginant su ne juodaodžiais pacientais, dėl klaidingų įsitikinimų apie biologinius skirtumus tarp juodaodžių ir baltaodžių, todėl juodaodžiams pacientams rečiau skiriami skausmą malšinantys vaistai, o jei skiriami – mažesnėmis dozėmis (Hoffman et al., 2016).

Atvejo analizė: rasiniai ir etniniai skirtumai tarp vidutinės gliukozės koncentracijos ir hemoglobino A1c rodiklio

A1C tyrimas matuoja vidutinį gliukozės (cukraus) kiekį kraujyje ir yra naudojamas prediabetui nustatyti arba padėti diagnozuoti 2 tipo diabetą. Tačiau A1C yra tik netiesioginis rodiklis ir nėra priežastiniu ryšiu susijęs su sveikatos rodikliais, nes egzistuoja daug veiksnių, galinčių tiesiogiai pakeisti ryšį tarp tiesioginių glikemijos matavimų (gliukozės koncentracijos kraujyje) ir A1C rodiklio. Be to, glikemijos ir A1C ryšys reikšmingai skiriasi tarp skirtingų asmenų ir net to paties asmens atžvilgiu skirtingais laikotarpiais. Taip pat tyrimuose nustatyta, kad afroamerikiečių pacientų hemoglobino A1C rodikliai yra reikšmingai aukštesni lyginant su baltaodžiais pacientas, kurių vidutinė gliukozės koncentracija yra tokia pati (Karter et al., 2023).

Jeigu DI sistema, skirta diabeto diagnostikai, yra mokoma naudoti A1C tyrimo rezultatus kaip pakaitinį glikemijos rodiklį, neatsižvelgiant į kitus veiksnius, pavyzdžiui,

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

paciento rasę, tai gali lemti per ankstyvą diabeto diagnozavimą ir netinkamą gydymą, o dėl to – šališką sveikatos priežiūros kokybę ir sveikatos netolygumus. Kaip pažymima (Alipour & Alipour, 2025) sisteminėje apžvalgoje apie šališkumus, galinčius paveikti DI ir mašininio mokymosi modelių teisingumą diabeto srityje (įskaitant matavimo šališkumą), apžvelgtuose tyrimuose aiškiai nurodoma, kad matavimo šališkumas gali persiduoti DI modeliams, jei nėra koreguojamas, tačiau nė viename iš jų tokie šališkumai modelių vystymo metu nebuvo sistemiškai įtraukti, aiškiai mažinami ar koreguojami atsižvelgiant į matavimo tikslumo skirtumus.

Agregavimo šališkumas

Agregavimo šališkumas atsiranda tada, kai vienas visiems tinkantis modelis taikomas duomenų rinkiniui, kuriame yra įvairių žmonių ar objektų grupių.

Galima nagrinėti pavyzdį, kai įvesties duomenys (pavyzdžiui, asmens pajamos) susiejami su juos aprašančiomis kategorijomis (pavyzdžiui, žemos, vidutinės, aukštos), darant prielaidą, kad toks priskyrimas yra nuoseklus visiems duomenų pogrupiams. Tačiau realybėje asmens socialinis ar kultūrinis kontekstas gali pakeisti tai, ką šie skaičiai iš tikrųjų reiškia. Pavyzdžiui, „aukštos“ pajamos mažame kaimo miestelyje ar mažų ar vidutinių pajamų valstybėje gali reikšti visai ką kita nei didmiestyje ar aukštų pajamų šalyje**.**

Atvejo analizė: skaitmeninės sveikatos priemonės pasyviai depresijos stebėsenai

Skaitmeninių priemonių naudojimas fiziologiniams ir elgsenos kintamiesiems matuoti pasyvios depresijos stebėsenos tikslais nagrinėjamas sisteminėje tematikos apžvalgoje (De Angel et al., 2022). Peržiūrėtuose straipsniuose buvo analizuojamos sąsajos tarp depresijos ir objektyvių elgsenos duomenų, gautų iš išmaniųjų telefonų ir dėvimų išmaniųjų įrenginių. Šie duomenys buvo paverčiami požymiais, kuriuos DI modeliai naudojo prognozėms sudaryti, atitinkančiais miegą, fizinį aktyvumą, cirkadinį ritmą, socialumą, buvimo vietą ir telefono naudojimą.

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

Tačiau autoriai pabrėžia nevienalytiškumą, atsirandantį dėl metodų, naudojamų šiems požymiams sukurti, įvairovės. Pavyzdžiui, požymis „miego kokybė“ gali būti apibrėžiamas matuojant prabudimų skaičių, bendrą budrumo minučių skaičių arba budėjimo ir miego santykį miego ciklo metu, taip pat būtina atsižvelgti į skirtumus, ką skirtingų įrenginių jutikliai priskiria „miegui“. Kadangi visi šie skirtumai dažnai nėra atskirai vertinami ir yra sujungiami į vieną bendrą kategoriją „miego kokybė“, o duomenų rinkinys gali būti sudarytas iš skirtingo socialinio, kultūrinio ar normatyvinio konteksto asmenų ar grupių, šis požymis skirtingoms grupėms ar individams gali turėti skirtingą reikšmę.

Tokių duomenų agregavimas į vieną bendrą kintamąjį gali lemti sistemą, kuri netinka nė vienai grupei arba suteikia pranašumą dominuojančiai populiacijai, jei kartu egzistuoja ir reprezentatyvumo šališkumas. Pavyzdžiui, yra duomenų, kad tarp vyrų ir moterų egzistuoja miego skirtumų, o moterys dažnai yra nepakankamai atstovaujamos miego tyrimuose. Be to, į miego modelius ir sutrikimus dažnai neįtraukiami kiti svarbūs veiksniai – neatskiriama socialinė lytis kaip socialinis konstruktas nuo biologinės lyties ir neatsižvelgiama į intersekcionalias tapatybes, apibrėžiamas pagal amžių, rasę ir socio-ekonominę klasę (Lok et al., 2024).

Mokymosi šališkumas

Mokymosi šališkumas atsiranda tada, kai modeliavimo sprendimai padidina veikimo skirtumus tarp skirtingų duomenų pavyzdžių ar grupių. Vienas pavyzdys susijęs su diferenciniu privatumu – mechanizmu, naudojamu DI sistemose siekiant užtikrinti, kad, analizuojant sistemos išvestį, nebūtų galima nustatyti, ar konkretaus asmens duomenys buvo įtraukti į pradinį duomenų rinkinį. Diferencinis privatumas naudojamas sveikatos priežiūros duomenų rinkiniuose siekiant apsaugoti jautrią pacientų informaciją, pavyzdžiui, retų ligų atvejais, kai kiekvieno paciento atvejis tam tikroje ligoninės aptarnaujamoje teritorijoje yra daugiau ar mažiau unikalus, todėl net ir nuasmeninus duomenis nėra labai sunku nustatyti asmens tapatybę. Tačiau nustatyta, kad diferencinis privatumas sumažina nepakankamai atstovaujamų duomenų įtaką

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

AEQUITAS

modeliui; todėl, jei DI sistema jau yra šališka, privatumo stiprinimo priemonės taikymas šį šališkumą dar labiau sustiprina (Bagdasaryan & Shmatikov, 2019).

Atvejo analizė: diferencinis privatumas ir sveikatos netolygumai

2018 m. rugsėjo mėn. JAV Gyventojų surašymo biuras paskelbė, kad duomenų rinkiniams, sudarytiems remiantis 2020 m. surašymo duomenimis, bus taikomas diferencinis privatumas. Tačiau (Santos-Lozada et al., 2020) tyrė, kaip diferencinio privatumo taikymas gali turėti įtakos informacijai apie mirtingumo netolygumus sveikatos srityje, ypač rasinių ar etninių mažumų grupėse mažose teritorijose ir mažiau urbanizuotose vietovėse. Jų rezultatai parodė, kad diferencinis privatumas stipriau paveiks ne ispanų kilmės juodaodžių ir ispanų kilmės gyventojų mirtingumo rodiklių įverčius lyginant su ispanų kilmės baltaodžių rodiklių įverčiais.

Šias išvadas patvirtino (Kurz et al., 2022), kurie parodė, kad taikant diferencinį privatumą tiems patiems duomenims gali būti iškraipomi naudojimosi Medicaid rodikliai, jau ir taip marginalizuotose rasinėse ir etninėse grupėse. Konkrečiai, tam tikrų apskrities, rasės ir etninės kilmės derinių naudojimosi rodikliai skyrėsi tarp diferencinio privatumo būdu apdorotų duomenų ir pradinių duomenų, o skirtumas kai kuriais atvejais viršijo 10 %. Be to, ne ispanų kilmės baltaodžiai buvo vienintelis rasinis ir etninis pogrupis, kuriam diferencinio privatumo algoritmas tiksliai atvaizdavo naudojimosi Medicaid rodiklius. Ši išvada gali turėti reikšmingų pasekmių sveikatos politikai, nes surašymo duomenys naudojami valstybės programų planavimui, išteklių paskirstymui, politikos vertinimui ir stebėsenai.

Vertinimo šališkumas

Vertinimo šališkumas atsiranda tada, kai konkrečiai užduočiai naudojami standartiniai duomenys, neatspindintys realios populiacijos. Standartinių duomenų rinkiniai yra standartizuoti rinkiniai, naudojami modelio kokybei matuoti ir sudarantys sąlygas kiekybiškai palyginti skirtingus modelius. Dėl to kyla rizika skatinti modelių vystymą ir diegimą, kurie gerai veikia tik toje duomenų dalyje, kuri atstovaujama standartiniame rinkinyje. Jei standartinis rinkinys pasižymi istoriniu, reprezentatyvumo ar matavimo

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

šališkumu, gali atsirasti diskriminacija pažeidžiamų pogrupių ar atskirų asmenų atžvilgiu.

Sveikatos priežiūros srityje, nepakankamas konkrečių populiacijų atstovavimas duomenų rinkiniuose, gali atsirasti dėl to, kad tam tikri asmenys ar grupės apskritai nepatenka į duomenų rinkinius (pavyzdžiui, nėščiosios dėl etinių apribojimų) arba dėl to, kad asmenys priskiriami netikslioms ar netinkamoms kategorijoms (pavyzdžiui, „mišri etninė kilmė“ ar „kita“). Pagrindinės to priežastys gali būti socialinės, techninės, teisinės ar etinės – struktūrinės kliūtys gauti sveikatos priežiūrą, techninės kliūtys rinkti ar skaitmeninti reikšmingus sveikatos duomenis, individualūs ir struktūriniai apribojimai dėl sutikimo dalytis duomenimis, teisiniai ar etiniai duomenų dalijimosi ribojimai, ribojantys duomenų prieinamumą, ir kiti veiksniai (Arora et al., 2023). Dėl to DI sistemos, mokytos tokiais standartiniais duomenimis, gali veikti prasčiau, kai jos yra taikomos nepakankamai atstovaujamų grupių asmenims. Tačiau svarbu pažymėti, kad standartinių duomenų tinkamumo problema yra bendresnio pobūdžio ir neapsiriboja vien šališkumo klausimu (Brooks, 2025).

Atvejo analizė: odos vaizdų duomenų rinkiniai

Odos vaizdų duomenų rinkiniuose tam tikros demografinės grupės yra nepakankamai atstovaujamos, nes dauguma šių rinkinių yra gaunami iš Šiaurės Amerikos ar Europos šalių populiacijų, tad daugiausia vaizduoja šviesios odos asmenis (Guo et al., 2021). Dėl didelių tokių duomenų rinkinių sudarymo kaštų ir sudėtingumo jie naudojami ne tik modelių mokymui, bet ir kaip standartiniai rinkiniai.

Atvejo analizė, iliustruojanti atsirandantį šališkumą – būtent odos vėžio vaizdų duomenų rinkiniai, naudojami prognozavimo modeliams mokyti – yra netinkamo standartinio rinkinio pavyzdys tais atvejais, kai naudotojų populiaciją sudaro nepakankamai atstovaujamos grupės (Guo et al., 2021). Panašus atvejis, nors ir nesusijęs su DI, parodo problemos universalumą. Jis buvo susijęs su pulso oksimetrais (prietaisais, matuojančiais kraujo prisotinimą deguonimi, naudojamais, pavyzdžiui,

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

ištikus širdies smūgiui ar širdies nepakankamumui), kurie, kaip paaiškėjo, daug tiksliau veikia esant šviesesnės pigmentacijos odai. (Sjoding et al., 2020).

Reprezentatyvumo, matavimo, agregavimo, mokymosi ir vertinimo šališkumai gali būti susiejami su techniniu šališkumu, kaip jį apibrėžė Friedman ir Nissenbaum (1996).

Diegimo šališkumas

Diegimo šališkumas atsiranda tada, kai nesutampa problema, kurią modelis yra skirtas spręsti, ir realus jo naudojimo būdas. Tai gali sukelti žalą, ypač kai kartu pasireiškia kognityviniai šališkumai, tokie kaip patvirtinimo šališkumas ir automatizavimo šališkumas. Diegimo šališkumas atitinka iškylantį šališkumą, kaip jį apibrėžė Friedman ir Nissenbaum (1996).

Atvejo analizė: domeno poslinkis

Duomenų poslinkio atvejis aprašytas iškylančio šališkumo poskyryje apie odos vėžio nustatymą. Papildomai galima apibrėžti domeno poslinkio atvejį – situaciją, kai sistema, gavusi reguliacinį leidimą yra įdiegta, ir naudojama klinikinėje praktikoje, tačiau taikoma kitokiai pacientų populiacijai nei ta, su kuria ji buvo mokyta. Pavyzdžiui, sistema gali būti sukurta ligoninei aukštų pajamų šalyje, o vėliau įdiegta žemų ar vidutinių pajamų šalyje, neatsižvelgiant į tokius veiksnius kaip pacientų sociodemografinės charakteristikos ar tai, ar jų bendras rizikos lygis sutampa su mokymo duomenyse buvusių pacientų rizikos lygiu (Vokinger et al., 2021).

Politikos pasekmės

D2.1 ataskaitoje susisteminti duomenys rodo, kad lyčių ir rasinis šališkumas biomedicininiame DI nėra atsitiktiniai ar pavieniai techniniai trūkumai, bet sisteminės rizikos, atsirandančios visame DI sistemų, naudojamų sveikatos priežiūroje, gyvavimo cikle. Širdies ir kraujagyslių ligų, depresijos ir diabeto srityse šališkumas kyla iš istoriškai iškreiptų klinikinių duomenų rinkinių, nelygių diagnostikos praktikų, pakaitinių

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

kintamųjų, įtvirtinančių struktūrines nelygybes, ir diegimo kontekstų, kuriuose nauda ir žala pasiskirsto netolygiai. Šios išvados patvirtina, kad biomedicininis DI tiesiogiai susijęs su keliomis Europos Sąjungos pagrindinių teisių chartijoje įtvirtintomis teisėmis ir principais – pirmiausia žmogaus orumu, lygybe prieš įstatymą ir diskriminacijos draudimu, taip pat teise į asmens neliečiamumą, teise į sveikatos priežiūrą, duomenų apsaugą ir teise į veiksmingą teisinę gynybą.

Atsižvelgiant į tai, ES ir nacionaliniai DI valdymo sveikatos priežiūros srityje politikos pagrindai turi šališkumo mažinimą laikyti ne savanorišku etiniu priedu, bet privalomu teisėto ir su teisėmis suderinamo DI diegimo elementu. Europos ir nacionalinės DI reguliavimo iniciatyvos sveikatos priežiūros srityje turi būti vertinamos platesniame DI pagrindinių teisių reguliavimo kontekste. ES DI aktas sukuria būtiną reguliacinį pagrindą, daugumą biomedicininių DI sistemų priskirdamas aukštos rizikos kategorijai, tačiau jo praktinis veiksmingumas priklausys nuo to, kaip pagrindinių teisių apsaugos priemonės bus įgyvendintos atitikties vertinimuose, naudojimo rinkoje stebėsenoje ir viešojo sektoriaus pirkimuose.

Pirma, prasmingos žmogaus priežiūros garantijos turi būti sustiprintos ir tiksliai apibrėžtos biomedicininių DI sistemų viso jų gyvavimo ciklo metu. Klinikiniai DI įrankiai, naudojami diagnostikai, rizikos vertinimui, patikrai ar gydymo sprendimų palaikymui, jokiu būdu neturi veikti kaip faktiškai autonomiški sprendimų priėmėjai. Žmogaus priežiūra turi apimti ne tik sveikatos priežiūros specialistų galimybę atmesti sistemos sprendimą, bet ir aiškią institucinę atsakomybę suprasti sistemos ribotumus, žinomas šališkumo rizikas ir veikimo skirtumus tarp pogrupių. Vadovaujantis Chartijoje įtvirtinta žmogaus orumo ir neliečiamumo apsauga, sveikatos priežiūros specialistai turi būti mokomi ir instituciškai remiami kritiškai vertinti DI rezultatus, o ne jais automatiškai pasikliauti. Tam būtina integruoti DI raštingumą, šališkumo suvokimą ir pagrindinių teisių mokymą į medicinos studijas ir nuolatinį profesinį tobulinimą.

Sveikatos priežiūros kontekste, skaidrumo pareigos turi būti aiškinamos plačiai. Pacientai ir sveikatos priežiūros paslaugų vartotojai turi būti informuoti kiekvieną kartą, kai DI sistemos naudojamos klinikiniam sprendimų priėmimui jų atžvilgiu,

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

įskaitant patikrą, prioritetų nustatymą ar rizikos vertinimą. Kai DI sugeneruoti rezultatai naudojami viešosiose sveikatos priežiūros paslaugose, jie turi būti aiškiai identifikuojami kaip tokie ir pateikiami kartu su suprantamais paaiškinimais apie jų vaidmenį, ribotumus ir žinomas šališkumo rizikas. Asmenys taip pat turi būti informuoti, kai jų asmens duomenys naudojami DI mokymui, testavimui ar nuolatiniam mokymuisi, ypač kai naudojami jautrūs sveikatos duomenys. Šios skaidrumo priemonės yra būtinos siekiant užtikrinti Chartijoje įtvirtintas teises į duomenų apsaugą ir veiksmingą teisinę gynybą bei sudaryti galimybę asmenims realiai ginčyti jiems nepalankius sprendimus.

Antra, poveikio vertinimas pagrindinėms teisėms turi tapti įprastu ir privalomu reikalavimu biomedicininėms DI sistemoms – jis turi apimti ne tik išankstinį vertinimą prieš pateikimą rinkai, bet ir nuolatinį vertinimą diegimo metu. D2.1 pateikti empiriniai duomenys rodo, kad daug šališkumo sukeliamos žalos paaiškėja tik tada, kai DI sistemos pradeda veikti realiose populiacijose ir klinikiniuose procesuose, ypač dėl intersekcionalių veiksnių, susijusių su lytimi, rase, amžiumi ir socio-ekonomine padėtimi. Todėl teisėmis grindžiami poveikio vertinimai turėtų būti privalomi aukštos rizikos medicininiam DI ir aiškiai vertinti skirtingų saugomų grupių veikimo skirtumus bei rezultatus, įtraukiant suinteresuotąsias šalis – pilietinės visuomenės organizacijas, pacientų atstovus ir lygybės institucijas.

Periodiniai biomedicininių DI sistemų auditai turėtų būti privalomi siekiant patikrinti nuolatinę atitiktį pagrindinių teisių standartams, ypatingą dėmesį skiriant šališkumui, duomenų rinkinių pokyčiams ir klinikinio naudojimo kaitai laikui bėgant. Nustačius nuolatinį ar neištaisomą diskriminacinį poveikį, turi egzistuoti aiškūs teisiniai ir instituciniai mechanizmai sistemos naudojimui apriboti, sustabdyti arba nutraukti. Teisė į sveikatos priežiūrą negali pateisinti DI priemonių naudojimo, jei jos sistemiškai blogina tam tikrų grupių padėtį, net jei bendrieji veikimo rodikliai atrodo geri.

Trečia, ES ir nacionalinės institucijos turi spręsti netinkamo DI panaudojimo ir antrinės žalos rizikas biomedicininio DI srityje, įskaitant kibernetinio saugumo pažeidžiamumus, galinčius pakenkti sistemų vientisumui ar sudaryti sąlygas piktybiškai manipuliuoti

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga

AEQUITAS

klinikiniais rezultatais, taip pat DI panaudojimą stebėsenai, profiliavimui ar atskirties praktikoms. Biomedicininės DI sistemos turi būti reguliariai tikrinamos saugumo požiūriu ir joms turi būti taikoma griežta incidentų pranešimo pareiga bei aiškūs atsakomybės mechanizmai.

Ketvirta, etinės ir atsakingos praktikos skatinimas turi būti įtvirtintas visoje biomedicininių DI sistemų vertės grandinėje. Vystytojai turi būti įpareigoti aktyviai mažinti šališkumo rizikas - rinkti reprezentatyvius duomenis, atsargiai parinkti tikslinius ir pakaitinius kintamuosius, atlikti validavimą pogrupių lygmeniu ir skaidriai pateikti veikimo rezultatus pagal lytį ir rasę. Viešųjų pirkimų ir finansavimo politika turi skatinti sprendimus, kurie įrodo patikimas ir nepriklausomai patikrintas šališkumo mažinimo praktikas.

Galiausiai, visuomenės atsparumo šališkam biomedicininiam DI stiprinimas reikalauja nuolatinių investicijų į visuomenės informavimą, pilietinės visuomenės įtraukimą ir tarpsektorinį bendradarbiavimą. Asmenys turi būti įgalinti suprasti savo teises DI pagrindu teikiamoje sveikatos priežiūroje ir žinoti bei suprasti esamus apsaugos mechanizmus. Pilietinės visuomenės organizacijos, lygybės institucijos ir pacientų grupės turi būti pripažįstamos esminiais dalyviais stebint DI poveikį ir formuojant politiką.

Apibendrinant, D2.1 ataskaitos išvados leidžia padaryti aiškią politinę išvadą: biomedicininis DI Europos Sąjungoje gali būti laikomas patikimu ir teisėtu tik tada, kai jo vystymas, diegimas ir valdymas yra tvirtai grindžiami pagrindinių teisių apsauga. ES DI aktas, aiškinamas per ES pagrindinių teisių chartijos prizmę ir įgyvendinamas per konkrečius priežiūros, poveikio vertinimo ir atskaitomybės mechanizmus, sudaro esminę galimybę užtikrinti, kad inovacijos sveikatos priežiūroje didintų lygybę, o ne palaikytų istorinius diskriminacijos modelius.

Finansuojama Europos Sąjungos lėšomis. Tačiau išreiškiamas požiūris ar nuomonė yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos švietimo ir kultūros vykdomosios įstaigos (EACEA) požiūrį ar nuomonę. Nei Europos Sąjunga, nei EACEA negali būti laikoma už juos atsakinga.Projekto kodas: 101215009 – AEQUITAS – CERV-2024-CHAR-LITI

Bendrai finansuoja Europos Sąjunga