

Tuukka Ruhanen

OMOP-CDM:N HYÖDYNTÄMINEN LÄÄKETIETEELLISESSÄ DATA-ANALYTIKASSA

Kandidaatintutkielma
Informaatioteknologian ja viestinnän tiedekunta
Tarkastajat: Mikko Nurminen
Joulukuu 2023

TIIVISTELMÄ

Tuukka Ruhanen: OMOP-CDM:n hyödyntäminen lääketieteellisessä data-analytiikassa
Kandidaatintutkielma
Tampereen yliopisto
Tieto- ja sähkötekniikan koulutusohjelma
Joulukuu 2023

Tässä kandidaatintutkielmassa tutkittiin Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)-tietomallin käyttöä lääketieteen ja terveydenhuollon data-analytiikassa. Päättävänä oli aiempien tutkimusten perusteella selvittää, millaisilla tavoilla OMOP-CDM:ää käytetään ja saavutetaanko sen käytöllä merkittävää hyötyä, joka voisi edesauttaa tietomallin laajempaa käyttöönottoa. Tietomallin laaja käyttö edistäisi terveydenhuollon pyrkimyksiä ymmärtää ja ratkaista niin yksilöiden kuin koko väestön terveystilanteita. Toisena päättävänä oli pohtia syitä vähäisen käytön taustalla eri näkökulmista.

Tutkielman aihe sijoittuu lääketieteen ja data-analytiikan rajapinnalle, minkä vuoksi katsauksessa tarkasteltavia julkaisuja haettiin kattavasti molempiin aloihin keskittyvistä tietokannoista, joista tärkeimpiä olivat PubMed, Springer ja ScienceDirect. Lähdejulkaisuja kerättiin vuosien 2018 ja 2023 väliltä, mikä osoittautui sopivaksi rajaukseksi, sillä vanhempia julkaisuja aiheesta ei juurikaan löytynyt. Tärkeimpänä valintakriteerinä julkaisuille oli OMOP-CDM:än merkittävyys tehdylle tutkimukselle tai tutkimuksen tuloksille. Aineistohaussa huomattiin, että OMOP-CDM mainittiin useissa julkaisuissa, mutta vain osassa sen käsittely oli oleellista työn kannalta. Varsinaisten tutkimusten rinnalla työssä käytettiin laajasti lähteitä selittämään käsitteitä, jotka ovat vahvasti sidoksissa teoriaan aiheen taustalla.

Tutkimuksen tulokset vahvistavat käsitystä, että OMOP-CDM on hyödyllinen väline tietoaaineistojen laajentamisessa. Potilastietojen standardoinnin myötä on mahdollista kasvattaa tietoaaineistoja yli paikallisten ja kansainvälisten rajojen, mikä mahdollistaa laajempien ja yleistettävien tutkimusten toteuttamisen verrattuna suppeampiin otoskooltaan rajoittuneisiin tutkimuksiin. Laajojen otoskokojen käyttö on olennaista korkealaatuisten kvantitatiivisten tutkimusten toteuttamisessa.

Työssä pohdittiin syitä tietomallin vähäiselle omaksumiselle. Heterogeenisen potilastiedon muuntaminen standardimuotoon on prosessi, joka vie paljon aikaa ja edellyttää yhteistyötä klinikoiden ja data-ammattilaisten välillä. Koska suurten työryhmien resursointi on kallista, monet organisaatiot eivät pidä tätä prioriteettina. Tämän ongelman ratkaisemiseksi pohdittiin mahdollisuutta potilastiedon rakenteellisempaan kirjaamiseen jo alusta alkaen, mikä nopeuttaisi myöhempää muuntamisprosessia, sekä työkalujen kehittämistä, joilla voitaisiin automatisoida muuntamisprosessia.

Avainsanat: OMOP-CDM, tietomalli, OHDSI, datatiede

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck -ohjelmalla.

SISÄLLYSLUETTELO

1. Johdanto	1
2. Tutkimusmenetelmä	3
3. Teoreettinen tausta	5
4. Datan standardointi ja datatiede terveydenhuollossa ja lääketieteellisessä tutkimuksessa.	7
4.1 Datan muuntaminen ja standardointi	7
4.2 Datatiede ja -analytiikka	8
5. Tapausesimerkkejä	11
5.1 Potilasennusteen luominen arkipäivän potilastiedosta	11
5.2 2. tyypin diabeetikoiden lääkehoito	12
5.3 Fenotyyppitystyökalu	12
6. Pohdinta	13
7. Yhteenveto	15
Lähteet	17

LYHENTEET JA MERKINNÄT

CDM	Common Data Model
EHR	Electronic Health Record
NLP	Natural Language Processing
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
RWD	Real-world data, arkipäivän potilastieto

1. JOHDANTO

Digitaalisen aikakauden myötä terveydenhuolto on kehittynyt nopeasti elektronisten terveystietojen (EHR) käyttöönoton myötä. Seurauksena datan rooli on yhä korostuneempi, kun pyrimme ymmärtämään ja vastaamaan sekä yksittäisten potilaiden että koko populaation terveysongelmiin. Tämän tutkielma antaa yleiskatsauksen kansainvälisestä standardoidusta tietomallista, Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM):ista. Katsaus käsittää tietomallin ominaisuudet, sen hyödyntämismahdollisuudet lääketieteen data-analytiikassa ja esimerkkitapauksia sen käytöstä. Aloite on askel kohti lääketieteellisen datan kansainvälistä standardointia, sillä se pyrkii standardoimaan kliinisen tiedon tallennuksen ja luo rajapintaa terveydenhuollon ja tietojenkäsittelytieteen välille [18]. Tutkielmassa käsitellään myös OHDSI (Observational Health Data Sciences and Informatics) -yhteisön kehittämiä avoimen lähdekoodin analytiikkatyökaluja. Nämä työkalut auttavat laajentamaan OMOP-CDM:n käyttöä datan analysoinnissa. OHDSI on avoin, vuonna 2014 alkunsa saanut yhteisö, jonka tavoitteena on edistää OMOP-CDM:n ja terveydenhuollon kehitystä. Valitsin OMOP-CDM:n tutkielmani aiheeksi, koska se on tullut tutuksi minulle työn kautta, ja jaan sen yhteisön kanssa samanlaiset arvot.

OMOP oli yhteistyöhanke yksityisen ja julkisen sektorin välillä, joka keskittyi havainnollisten lääketieteellisten tietokantojen tehokkaaseen käyttöön lääketuotteiden vaikutusten tutkimisessa. Projekti kesti viisi vuotta ja saavutti tavoitteensa: se teki metodologista tutkimusta, loi analysointityökaluja erityyppisille datalähteille ja perusti yhteisresurssin, jota kansainvälinen tutkimusyhteisö voi hyödyntää. [14] Tutkimustuloksia on esitelty useissa tieteellisissä konferensseissa ja ne ovat saaneet julkaisuhuomiota. OMOP-CDM on edelleen käytössä ja tukee monenlaista tutkimustyötä. OMOP-CDM on avoimen lähdekoodin standardi ja saatavilla julkisesti.

Terveydenhuollon data-analyysillä on keskeinen rooli potilaiden hoidon tehostamisessa, resurssien hallinnassa sekä kliinisen ja hallinnollisen päätöksenteon tukemisessa. Data-analyysin avulla voidaan esimerkiksi ennustaa tautien leviämistä tai sairauksien puhkeamista, luoda yksilöllisiä hoitosuunnitelmia ja valvoa julkisen terveyden tilaa. [2] Potilastyön lisäksi se mahdollistaa kustannustehokkuuden parantamisen, sillä kattavammat analyysit johtavat resurssien tarkempaan kohdentamiseen ja vähentävät yli diagnostiikkaa ja turhia toimenpiteitä.

Terveydenhuollosta syntynyt data on usein saatavilla vain paikallisesti, jopa valtioiden rajojen sisäpuolella. Suomessa on huomattu, että potilastieto siirtyy heikosti potilastietojärjestelmien ja hyvinvointialueiden välillä, mistä syystä standardisoitu data on ehdoton edellytys kansainvälisen hajautetun analyysin mahdollistamiseksi. [10] Tästä syystä tutkielman näkökulmaksi valikoitui kansainvälinen tutkimuskenttä. Työn suunnitteluvaiheessa tarkoituksena oli kirjoittaa tutkielma Suomen näkökulmasta, mutta suomalainen FinOMOP-hanke on tuore, eikä se ole vielä kerryttänyt vielä riittävästi tieteellisiä julkaisuja.

Tutkielma vastaa seuraavaan tutkimuskysymykseen: Kuinka OMOP-CDM:ää hyödynnetään lääketieteellisessä tutkimuksessa ja terveydenhuollossa? Kysymykseen pyritään vastaamaan esittelemällä monipuolisesti julkaisuja, joissa OMOP-CDM on ollut keskeisessä roolissa. Tämän kirjallisuuskatsauksen tavoitteena on lisätä ymmärrystä OMOP-CDM:n käytöstä, koota yhteen sen avulla saavutettuja onnistumisia, ja tutkia, mikä tekee tästä tietomallista merkittävän. Analysoimalla mallin käyttötapauksia ja saavutettuja tuloksia voimme hahmottaa, miten OMOP-CDM voi edistää lääketieteen kehitystä ja tukea päätöksentekoa terveydenhuollossa.

Tutkielmassa tarkastellaan ensiksi käytettyä tutkimusmenetelmää ja metodeita. Tutkimusmenetelmän esittelyn jälkeen avataan työn kannalta keskeisiä käsitteitä, ja tarjotaan näin lukijalle hyvä perustan työn ymmärtämiseen. Tutkielman ydinosassa keskitytään OMOP-CDM:n hyödyntämiseen konkreettisten tapausesimerkkien kautta eri tutkimuksissa ja julkaisuissa. Ennen näiden tapausesimerkkien tarkastelua lukija tutustutetaan data-analyysin perusteisiin sekä OHDSI:n (Observational Health Data Sciences and Informatics) tarjoamiin analytiikkatyökaluihin. Tutkielman loppuosassa analysoidaan ja keskustellaan havaituista löydöksistä, pohditaan niiden merkitystä ja tehdään yhteenveto tutkielman keskeisistä johtopäätöksistä ja havainnoista.

2. TUTKIMUSMENETELMÄ

Tässä luvussa kuvataan yksityiskohtaisesti kirjallisuuskatsauksen suorittamiseen käytetyt menetelmät, mukaan lukien tietokannat, hakutermit, aikarajaukset ja arviointikriteerit. Menetelmien kuvaus pyrkii tarjoamaan kattavan ja läpinäkyvän yleiskuvan tutkimusprosessista, jotta tulokset olisivat toistettavissa ja arvioitavissa.

Tämä kirjallisuuskatsaus pohjautuu pääosin vertaisarvioituihin tieteellisiin artikkeleihin, joiden tueksi käytetään lehtiartikkeleita, teknistä dokumentaatiota ja avoimen yhteisön ylläpitämää data- ja työkalupankkia. Tutkielman tulee määriteltyjen minimivaatimuksien mukaan sisältää 10-15 vertaisarvioitua artikkelia, minkä tarkoituksena on rajoittaa tutkielma sopivan mittaiseksi kandidaatintyön vaatimiin tavoitteisiin.

Suurin osa vertaisarvioituista artikkeleista on etsitty ProQuest-, PubMed-, Springer- ja ScienceDirect-tietokannoista. Näistä tietokannoista on löytynyt kattavimmin omaa aihetta ni sivuavia artikkeleita kuin puhtaasti tietojenkäsittelyaiheisiin keskittyvistä tietokannoista. Kirjallisuuskatsauksen ajankohtaisuuden varmistamiseksi artikkelit on rajattu viiden vuoden aikavälille, eli artikkelit pyrittiin valitsemaan vuosien 2019-2023 väliltä. Taulukossa 2.1 on esitelty parhaita hakutermejä ja osumia, joita niillä on saatu kyseisistä tietokannoista. Suurin osa tutkielmassa käytetyistä tieteellisistä lähteistä löydettiin näistä tietokannoista. Muutama sopiva lähde löydettiin seuraamalla artikkeleiden lähteitä.

Taulukko 2.1. Käytettyjä hakutermejä

Tietokanta	Hakutermit	Löydetyt julkaisut
ProQuest	noft(OMOP) AND noft(EHR)	16
PubMed	"OMOP-CDM"AND OMOP	122
ScienceDirect	OMOP AND "data analysis"	121
TUNI Andor	OMOP AND "data analysis"	32

Tutkielman aiheen luonteen vuoksi hakutermit ovat usein akronyymejä, mikä aiheutti hakutuloksiin virheellisiä osumia. Virheelliset osumat olivat esimerkiksi artikkeleita, joissa on kerran mainittu akronyyni OMOP-CDM ilman, että siihen on sen enempää otettu kantaa tai käsitelty. Tämän kaltaisten osumien välttämiseksi, taulukossa 2.1 esitelty hakutermit "noft(OMOP) AND noft(EHR)", onnistui hyvin. Hakumääreellä "noft", haku määritellään

suoritettavaksi abstrakteihin ja avainsanoihin varsinaisen julkaisun käsittelytekstin sijasta.

Lähteiden laadun arviointiin käytettiin Julkaisufoorumin Jufo-portaalia. Jufo-portaalista voidaan hakea ja vertailla erilaisten julkaisukanavien tasoluokituksia. Tässä työssä kaikki vertaisarviodut lähdeartikkelit on valittu vähintään tasoluokituksen 1 saaneista julkaisumedioista. Tasoluokituksen varmistamisen jälkeen, ennen artikkelin valintaa lähteeksi, artikkelista käytiin läpi abstrakti-, johdanto- ja yhteenvetoluvut. Jos näissä luvuissa tutkielman aihetta sivuttiin riittävästi tai suoranaisesti käsiteltiin, julkaisu otettiin talteen työn rajaamista varten.

3. TOOREETTINEN TAUSTA

Tässä luvussa taustoitetaan käsitteitä ja teoriaa lääketieteellisistä tietomalleista ja data-analytiikasta. Tarkoitus on varmistaa, että tutkielmassa käytetyt käsitteet on määritelty selkeästi ja työ tarjoaa riittävästi tietoa tietomallien taustalla olevasta teoriasta. Työn suppeuden vuoksi tämän luvun ei ole tarkoitus olla tarkka kuvaus teoriasta, vaan antaa korkeamman tason käsitys tutkielman aiheesta. Tarjoamalla yleiskatsauksen keskeisiin termeihin, kuten tietomalleihin, lääketieteelliseen dataan, OMOP-CDM:ään ja data-analytiikkaan, on olennaista monimutkaisempien tutkimusten ja keskusteluiden ymmärtämiseksi. Myöhemmissä luvuissa käsitellään OMOP-CDM:n kansainvälisiä sovelluksia, etuja ja haasteita terveydenhuollon data-analytiikassa.

Tietomalli on abstrakti malli, joka järjestää tietoalkioita ja määrittelee, miten ne vaikuttavat toisiinsa. Koska nämä tietoalkiot kuvaavat usein reaali maailman ilmiöitä, voidaan tietomallin sanoa mallintavan todellisuutta. [22] Esimerkkinä tietomalli voi pitää sisällään dataa olemassa olevien ihmisten lääketieteellisestä historiasta, jolloin tietomalli on eräänlainen kuva reaali maailmasta. Tietomalleilla on keskeinen rooli tietojärjestelmien välisessä kommunikaatiossa: ne toimivat välineenä, joka auttaa määrittelemään datan rakenteen ja standardit [22]. Tietomalleja luodaan ja havainnollistetaan yleensä käyttäen graafista notaatiota, ja ne saatetaan tietokoneohjelmoinnissa tuntea myös tietorakenteina.

Terveydenhuollon data-analytiikassa tietomallit ovat muutakin kuin pelkkiä tietoalkioiden ja niiden suhteiden muodollisia esityksiä. Ne ovat tärkein komponentti sähköisten potilastietojen rakenteistamisessa ja hallinnassa. Tietomallit toimivat pohjana sille, miten erityyppisiä terveydenhuollon tietoja, kuten potilastietoja, lääketieteellisiä toimenpiteitä ja lääkemääräyksiä, sovitetaan toisiinsa. Lääketieteellisten tietomallien kokonaismäärä on hyvin suuri lääketieteellisen terminologian monimutkaisuuden vuoksi, mistä syystä kerätty data on saavutettavissa yleensä vain hyvin paikallisesti. [4] Datan standardointi on tärkeää, kun halutaan varmistaa tietojen yhteensopivuus erilaisten terveydenhuoltojärjestelmien kesken, ja mahdollistaa laajamittainen lääketieteellinen tutkimus. Tämä ajatus oli koko OMOP-CDM-hankkeen taustalla.

OMOP-CDM on poikkeuksellinen tietomalli, joka on suunniteltu terveydenhuollon data-analytiikkaa varten. Se erottuu monista muista tietomalleista ominaisuuksillaan yhdistää eri lähteistä tulevat havainnolliset potilastiedot yhteiseen, analysoitavaan formaattiin. [14]

Lähteet voivat sijaita missä päin maailmaa tahansa ja OMOP-CDM:n tarkoituksena on yhdistää potilastiedot yli rajojen, jotka yleensä lamaavat kansainvälistä tutkimusyhteistyötä. Toisin kuin monet perinteiset tietomallit, jotka on suunniteltu pääosin organisaatioiden sisäisiin tarpeisiin, OMOP-CDM on luotu tukemaan laajamittaisia tutkimuksia, jotka sisältävät dataa useista eri järjestelmistä ja maista. Sen avulla voidaan tehdä vertailukelpoisia ja toistettavissa olevia analyyskejä, jotka keskittyvät erilaisten hoitojen vaikutuksiin ja tuloksiin. Tämä tietomalli myös noudattaa erityisiä suunnitteluperiaatteita, kuten tietosuojaa ja skaalautuvuutta, mikä tekee siitä soveltuvan moniin erilaisiin tutkimus- ja analysointitarpeisiin. Lisäksi, toisin kuin monet muut mallit, OMOP-CDM on teknologianeutraali eli sen voi toteuttaa missä tahansa relaatiotietokannassa. [14]

Lääketieteellinen data ei koostu pelkästään rakenteisesta datasta, vaan se on luonteeltaan heterogeenistä. Lääketieteellinen data muodostuu sähköisistä potilastiedoista, joihin kuuluvat lääkitykset, potilastekstit ja laboratoriomittaukset. Siihen kuuluvat myös lääketieteellisestä kuvantamisesta syntyvä data, sensoridata, biolääketieteellinen signaalianalyysi kuten sydänfilmit (EKG) ja aivosähkökäyrät (EEG), genomidata, kliiniset tekstit, lääketieteellinen kirjallisuus ja sosiaalinen media [16]. Moniulotteisen ja heterogeenisen datan kokonaisvaltaiseen hyödyntämiseen tarvitaan datan harmonisaatiota ja standardeja, sillä data-analytiikka on parhaimmillaan, kun analysoitava data on yhtenevää.

Data-analytiikka on prosessi, joka tulkitsee kvantitatiivista dataa tuottaakseen laadullista ymmärrystä, vastataakseen kysymyksiin ja tunnistaa trendejä. Keskeisiä analyysityyppejä ovat deskriptiivinen analyysi, joka tarkastelee ja kuvaa jo tapahtunutta; diagnostinen analyysi, joka pyrkii ymmärtämään tapahtuman syyt; prediktiivinen analyysi, joka tutkii historiallista dataa ja aiempia trendejä ennustaakseen tulevaa; sekä preskriptiivinen analyysi, joka tunnistaa toimenpiteet, joita yksilö tai organisaatio voi toteuttaa saavuttaakseen tulevaisuuden tavoitteita [2]. Nämä analyysimenetelmät mahdollistavat muun muassa virusten tarttuvuuden selvittämisen, diagnoosien laatimisen oireiden perusteella, kausittaisten sairauksien leviämisen ennustamisen ja ennaltaehkäisevien hoitosuunnitelmien laatimisen [2]. Data-analytiikan asema lääketieteessä vahvistuu jatkuvasti, kun ihmisten väkiluku kasvaa ja lääketieteellisen datan määrä sen mukana.

4. DATAN STANDARDOINTI JA DATATIEDE TERVEYDENHUOLLOSSA JA LÄÄKETIETEELLISESSÄ TUTKIMUKSESSA

Tässä luvussa esitellään lähdedatan muuntamista yhteensopivaksi OMOP-CDM:n kanssa, datatiedettä ja muutamia yleisiä data-analyysimenetelmiä, joita käytetään nykyaikana lääketieteen tutkimuksessa ja terveydenhuollossa. Näihin kuuluvat esimerkiksi tilastollinen analyysi, koneoppiminen ja päätöksenteon tuki.

4.1 Datan muuntaminen ja standardointi

Terveydenhuollossa käytetään useita erilaisia tietojärjestelmiä useilta eri toimittajilta. Kuntaliiton vuonna 2020 luomasta raportista [11] käy ilmi, että Suomessa oli vuonna 2020 käytössä 11 potilastietojärjestelmää kaiken kaikkiaan yhdeksältä eri toimittajalta. Suurin osa näistä potilastietojärjestelmistä ei ole keskenään yhteensopivia. Tämä johtaa siihen, että potilastiedot ovat hajallaan useissa järjestelmissä, mikä vaikeuttaa niiden tehokasta hyödyntämistä sekä laadullisessa tutkimuksessa että potilastyössä.

Kliinikoiden työn kannalta tiedon tallentaminen standardimuotoon ei välttämättä ole käytännöllistä, mutta tutkimuskäyttöä ajatellen se olisi optimaalista. Kun käytössä on useita potilastietojärjestelmiä, joihin tallennettua dataa halutaan standardisoida, tarvitaan yhteistä tietomallia. OMOP-CDM:n ideana on toimia rajapintana kliinisen työn ja tutkimuksen välillä. Onnistuneesti toteutetun tiedon standardisoinnin myötä potilastietojen hyödyntäminen ei rajoittuisi vain yhteen suljettuun tietojärjestelmään. [10]

OMOP-CDM pohjautuu lääketieteelliseen sanastoon, joka on luotu kansainvälisessä yhteistyössä. Sanasto on rakennettu yksityiskohtaisen kartoitusprosessin kautta standardoimaan erilaisia terveydenhuollon käsitteitä data-analyysin ja yhteensopivuuden mahdollistamiseksi. Kartoitusprosessissa yhdistetään paikallisen lähdejärjestelmän käsite ja koodijärjestelmän koodi. Käsitteitä voivat olla esimerkiksi diagnoosit, oireet, mittaustulokset, lääkitykset tai toimenpiteet. [7] Kuvassa 4.1 on esitelty kaksi esimerkkitapausta samankaltaisten käsitteiden kartoittamisesta. Kuva havainnollistaa sitä, kuinka lähdejärjestelmän käsite yhdistetään standardissa määriteltyyn koodiin ja käsitteeseen. Huomataan myös, että "Trigger finger"-käsitteellä on sama nimi standardijärjestelmässä, mutta

lähdekäsite "Amebiasis" on standardijärjestelmässä "Amebic infection".

Single "Maps to" for mapping to an Equivalent Standard Concept

Source Concept Code	Source Concept Name	Relationship ID	Standard Concept ID	Standard Concept Code	Standard Concept Name
M65.3	Trigger finger	Maps to	763891	448251000124102	Trigger finger
A06	Amebiasis	Maps to	438959	111910009	Amebic infection

Kuva 4.1. Esimerkki samankaltaisten käsitteiden kartoituksesta [7]

Kartoitusprosessin tueksi OHDSI on luonut Usagi-työkalun, joka ehdottaa kartoituksia koodikuvausten tekstuaalisen samankaltaisuuden perusteella. Käyttäjän työ helpottuu, kun voidaan vain tarkistaa ja hyväksyä Usagin tekemät ehdotukset. Erityisesti vieraskielisen lähdekoodiston kanssa tästä voi olla käyttäjälle apua. Usagi osaa toisaalta ehdottaa vain käsitteitä, jotka on jo määritelty standardikonsepteiksi sanastossa. Tämä vaikeuttaa erityisesti työskentelyä muilla kielillä kuin englanniksi, joten Usagi antaa käyttäjälle mahdollisuuden etsiä sopivaa konseptia myös manuaalisesti sanastosta [13, 20]. Usagilla on myös tutkimuskäyttöä, esimerkiksi Liu ja muut [12] käyttivät sitä käsitteiden normalisointiin tutkimuksessaan, jossa he tutkivat ontologian käsitteiden käytön tehokkuutta kliinisten tutkimusten luokittelussa.

Natiividatan muuntaminen OMOP-CDM standardiksi vaatii ETL (Extract, Transform, Load) -prosessia. ETL-prosessissa data järjestetään uudelleen CDM:ään sopivaksi yleensä automatisoidusti käyttämällä esimerkiksi SQL-skriptejä. OHDSIN kirjan [20] kappaleessa 6 ETL-prosessin vaiheet esitellään seuraavasti: suunnittelu data- ja CDM-asiantuntijoiden toimesta, koodikartoituksen luominen lääketieteen asiantuntijoiden toimesta, sekä ETL:n tekninen toteutus ja laadunvalvonta, johon osallistuvat kaikki edellämainitut.

OHDSI on kehittänyt avoimen lähdekoodin työkaluja, kuten WhiteRabbit ja Rabbit-In-A-Hat, joita voidaan käyttää ETL:n suunnittelun tukena. WhiteRabbit-työkalu skannaa datan ja luo raportin, joka sisältää suunnittelulle oleellisia tietoja taulukoista, kentistä ja arvoista. Rabbit-In-A-Hat toimii yhdessä WhiteRabbitin skannauksen kanssa suunnittelemalla ETL-prosessia. Se ei varsinaisesti luo koodia, vaan keskittyy taulukoiden ja kenttien väliin kartoitukseen. [13]

4.2 Datatiede ja -analytiikka

Nykyaikaisen laskentatehon ansiosta suurten datamäärien analysointi onnistuu nopeasti, mikä on laajentanut mahdollisuuksia hyödyntää lääketieteellistä dataa entistä tehokkaammin ja kattavammin. Tässä osiossa käsitellään datatieteen roolia lääketieteen tukena, esitellen yleisimpiä analyysimenetelmiä, joita hyödyntämällä voidaan saavuttaa merkittäviä etuja nykyaikaisessa terveydenhuollossa.

Datatiede on nopeasti kasvava poikkitieteellinen ala, joka on keskeisessä asemassa nykyaikaisessa lääketieteessä ja terveydenhuollossa. Se on mahdollistanut suurten ja monimutkaisten tietoaisteiden, kuten potilaiden demografisten tietojen, hoitotulosten ja sähköisten potilastietojen, tehokkaan hallinnan, analysoinnin ja yhdistämisen [19]. Subrahmanya ym. korostavat lääketieteellisen datatieteen keskeisiä etuja, jotka liittyvät koneoppimisen algoritmien, tiedonlouhinnan ja suurten tietomäärien analytiikan käyttöön. Nämä työkalut auttavat tuomaan esiin laajojen tietoaisteiden monimutkaisia yhteyksiä, parantamaan päätöksentekoa ja kehittämään potilashoitoa. Datatiede edesauttaa sairauksien varhaista havaitsemista, yksilöllistettyä lääkitystä ja potilaskeskeisiä hoitomuotoja, sekä eri tietolähteiden integroimista kokonaisvaltaisiin terveystietoihin [19].

Tilastolliset menetelmät ovat olennaisia terveydenhuollon datan analyysissä, kliinisessä päätöksenteossa ja lääketieteellisessä tutkimuksessa. Määrällisen datan analyysissä kuvaileva tilastoanalyysi ja tilastollinen päättely erotellaan tyypillisesti toisistaan. Kuvaavat tilastot (descriptive statistics) ovat olennainen osa määrällisen datan analysointia, sillä ne tarjoavat tietoa datan tyypillisistä arvoista ja niiden vaihtelusta. Niissä käytetään keskiluvun mittareita, kuten keskiarvo, mediaani ja moodi, sekä vaihtelun mittareita, kuten varianssi, keskihajonta ja vaihteluväli, datan ominaisuuksien ymmärtämiseksi. [6]

Päättelytilastot (inferential statistics) laajentavat analyysin ulottuvuutta pelkästä datan kuvailemisesta, mahdollistaen johtopäätösten tekemisen datasta. Tärkeitä päättelytilastollisia menetelmiä ovat muun muassa t-testit ja varianssianalyysi (ANOVA), jotka vertailevat ryhmien keskiarvoja määrittääkseen niiden tilastollisen merkittävyyden. T-testit tarkastelevat kahden ryhmän keskiarvojen eroja, kun taas ANOVA soveltuu useamman ryhmän keskiarvojen vertailuun. Päättelytilastot sisältävät myös korrelaatio- ja regressioanalyysit, jotka tutkivat muuttujien välistä suhdetta ja ennustavat yhden muuttujan arvoa toisen perusteella. [6] Ennakoivan analytiikan osana toimivat koneoppimisen algoritmit ovat nousseet merkittävämpään rooliin nykyaikana.

Koneoppimisalgoritmien käyttö klinisen työn tukena on lisääntynyt. Erityisesti syväoppimisalgoritmit ovat korostaneet asemaansa diagnostiikassa tehostamalla lääketieteen sovelluksia, kuten kuvantamisen diagnostiikkaa ja kliinistä genomitutkimusta. Yhdysvaltain elintarvike ja -lääkeviraston (FDA) hyväksymänä konenäöllä on saavutettu teknologisia harppauksia lääketieteellisen kuvantamisen ja patologian saralla. Konenäköä hyödynnetään erityisesti magneettisen resonanssikuvantamisen (MRI) ja patologisten näytteiden analysoinnissa. [3] Modernin lääketieteen kuvantamisteknologian tukena on järkevää hyödyntää konenäköä. Dias ja Torkamani nostavat artikkelissaan [3] yhtenä sovellusalueena esiin myös genomitutkimuksessa fenotyyppisen tiedon louhinnan, jota he pitävä lupaavana koneoppimisen sovelluksena.

Yksi merkittävä data-analytiikan tyyppi, erityisesti terveydenhuollossa, on päätöksenteon tuki, joka koostuu pääasiassa deskriptiivisestä ja/tai prediktiivisestä analyysistä. Islamin

ja muiden mukaan päätöksenteon tuen analyysi kohdistuu lähinnä sydän- ja verisuonisairauksiin, syöpään, diabetekseen ja akuuttihoitoon potilaisiin. [8] Päätöksenteon tuella on mahdollista parantaa potilasturvallisuutta keventämällä kliinikoiden työtaakkaa ja vähentämällä inhimillisten virheiden esiintyvyyttä.

5. TAPAUSESIMERKKEJÄ

Tässä luvussa tutustutaan muutamaaan tapausesimerkkiin, joissa on hyödynnetty OMOP-CDM -tietomallia ja OHDSI:n kehittämiä työkaluja. Esimerkit on valikoitu sen perusteella, kuinka keskeisessä roolissa OMOP-CDM:n käyttö on ollut, ja miten se on vaikuttanut julkaisujen sisältöön.

5.1 Potilasennusteen luominen arkipäivän potilastiedosta

Johnstonin ym. [9] tutkimuksessa todettiin, että lihavuusleikkauspotilaiden vaste toimenpiteelle vaihtelee. Tutkimusryhmä loi patient-level prediction (PLP, potilastason ennuste) -mallin, jolla voitiin ennustaa lihavuusleikatun tyyppin 2 diabeetikon lääkehoidon tarvetta toimenpiteen jälkeen. Tutkimukseen otettiin yli 16 000 potilasta kahdesta eri terveydenhuoltolaitoksesta Yhdysvalloissa. Näistä potilaista kartoitettiin erilaisia mahdollisesti lopputulemaan vaikuttavia tekijöitä, kuten aiemmat diagnoosit, laboratoriolöydökset ja lääkeytykset, sekä analysoitiin, ketkä pystyivät lopettamaan diabeteslääkkeet lihavuusleikkauksen jälkeen. Näiden perusteella luotiin malli, joka ennusti potilaan ennakkotietojen perusteella, kuinka todennäköisesti hän pääsisi lopettamaan diabeteslääkkeet, jos hänelle tehtäisiin lihavuusleikkaus.

Tutkimusryhmä päätti hyödyntää OMOP-CDM:ää erityisesti sen avoimen lähdekoodin vuoksi. Valintaan vaikuttivat myös OHDSI:n luomat apuohjelmat sanaston luomista varten. OMOP:in avulla PLP-malli voitiin luoda tutkimukseen osallistuneiden terveydenhuoltoyksiköiden tietokantaan, mutta yleistää sitten käytettäväksi missä tahansa OMOP-CDM:ää käyttävässä yksikössä. Tutkimuksessa käytettiin OHDSI:n standardisanastoa, ja heidän oli näin ollen mahdollista kartoittaa aineisto suhteellisen automaattisesti, mikä oli luonnollisesti tehokasta manuaaliseen kartoittamiseen verrattuna. Näin tutkimukseen pystyttiin sisällyttämään myös sellaisia tekijöitä, joiden ei etukäteen tiedetty vaikuttavan potilaan lopputulokseen. OMOP-CDM:n hyötynä ryhmä mainitsee vielä tehokkaan ja helpon toisintamismahdollisuuden muissa tutkimusryhmissä, mikä pienentää virheiden mahdollisuutta tutkimuksen jäljentämisessä. Tutkimusryhmä kannustaa jatkokehityksessä luomaan yhä uusia vastaavia työkaluja, joita voidaan hyödyntää tosielämän päätöksenteossa lääketieteessä, tässä tapauksessa sen päättämisessä, kannattaako potilaalle tehdä lihavuusleikkaus. On toisaalta oleellista tiedostaa, että kyseessä on edelleen vain pää-

töksentekoa tukeva aputyökalu; kyseisellä PLP-ohjelmalla saadaan todennäköisyys lääkityksen tarpeelle toimenpiteen jälkeen, ja edelleen jää kliinikon vastuulle arvioida, mikä on leikkauskynnyksen ylittävä todennäköisyys. [9]

5.2 2. tyypin diabeetikoiden lääkehoito

Vashishtin ym. [21] tutkimus on puolestaan esimerkki OMOP-CDM:n mahdollistamasta suuren potilasaineiston kliinisestä tutkimuksesta. Tutkimuksessa vertailtiin tyypin 2 diabeteksen lääkehoitovaihtoehtoja keskenään. Vaikka aiheesta on tehty viime vuosikymmeninä paljon lääketieteellistä ja farmaseuttista tutkimusta, on tutkimusnäyttö ollut keskenään ristiriitaista. Tässä tutkimuksessa yhdistettiin useiden aikaisempien tutkimusten potilastiedot OMOP-CDM:n avulla yhdeksi aineistoksi, jolloin saatiin lähes neljännesmiljardin potilaan kansainvälinen aineisto. OMOP-CDM:n avulla saatiin siis yhdistettyä kahdeksan eri yksikön potilastiedot kolmesta eri maasta. Aineistosta kartoitettiin potilaiden ikä, sukupuoli, diagnoosit, lääkitykset ja tehdyt toimenpiteet. Edes näin suuressa aineistossa eri lääkeaineryhmien välille ei muodostunut merkittävää eroa. Tutkimusryhmä pohti, että tutkimus kannattaa myöhemmin uusia ja ottaa mukaan vielä useampia OMOP:ia hyödyntäviä laitoksia. Lisäksi tutkimusryhmä kehotti hyödyntämään OHDSI:n työkaluja myös jatkossa suuren otoskoon tutkimusten tekemisessä.

5.3 Fenotyyppitystyökalu

Sharma ym. [17] kehittivät OMOP-CDM:ää hyödyntävän fenotyyppitystyökalun. Fenotyyppityksellä tarkoitettiin potilaiden luokittelua heidän ominaisuuksiensa perusteella, tässä tutkimuksessa lihavuuden eri liitännäissairauksien mukaan. Työkalulla saatiin kerättyä kliinisesti oleelliset tiedot paitsi järjestetystä, myös järjestämättömästä datasta eli potilaskertomuksista. Järjestämättömän datan keräämisessä hyödynnettiin NLP-järjestelmää, ja tiedot tallennettiin OMOP-CDM -formaattiin. Työkalun keskeinen tarkoitus on nimenomaan mahdollistaa myöhemmin eri laitosten ja tietojärjestelmien datan yhtenäinen käsittely, jolloin voidaan tehdä jatkotutkimusta suuremmilla potilasaineistoilla. Lisäksi tarkoituksena oli toimia suunnannäyttäjänä fenotyyppitysmallin luomisessa, niin että toiset tutkimusryhmät voisivat hyödyntää sitä myös muissa yhteyksissä kuin tämän esimerkin lihavuuden tutkimisessa. Tutkimusryhmän mukaan OMOP-CDM:n käyttämisen etuna ovat erityisesti sen tarjoamat tietokannan hallintajärjestelmät.

6. POHDINTA

Tämä kandidaatintutkielma keskittyy OMOP-CDM:n rooliin lääketieteellisessä data-analytiikassa ja tarjoaa yleiskatsauksen sen käytöstä, sovelluksista ja mahdollisuuksista. Tässä osiossa syvennytään pohtimaan ja arvioimaan niitä tekijöitä ja haasteita, jotka vaikuttavat OMOP-CDM:n laajamittaiseen käyttöönottoon ja soveltuvuuteen sekä kliinisessä että tutkimusympäristössä. Lisäksi pohditaan tietomallin mahdollisuuksia tulevaisuudessa, sen laajentamista uusille erikoisaloille sekä niitä teknologisia ja eettisiä näkökohtia, jotka ovat keskeisiä sen onnistuneessa toteutuksessa. Lisäksi tämä osio tarjoaa kriittisen näkökulman OMOP-CDM:n hyödyntämiseen liittyviin haasteisiin ja mahdollisuuksiin korostaen alueita, joilla lisätutkimus ja kehitys ovat tarpeen.

Viime vuosina OMOP-CDM -tietomalli on vakiintunut laajalti hyväksytyksi standardiksi sairaalatietojen yhtenäistämiseen sekä Euroopassa että Yhdysvalloissa. Tämän muutoksen taustalla on UK Health Data Research Alliancen suositus, jossa OMOP-CDM -mallia ehdotetaan kaikille yrityksille ja organisaatioille, jotka ovat harkitsemassa tutkimustietojen standardien käyttöönottoa. [1] EHDEN (European Health Data & Evidence Network) raportoi vuonna 2022, että heillä on 187 datakumppania 29 eri maasta, jotka kartoittavat potilasdataa OMOP:in yhteiseen tietomalliin [5]. Määrä on noussut Laitisen [10] vuonna 2021 ilmoittamasta 143:sta datakumppanista 27:stä ei maasta. Datakumppanien määrän nousu ennustaa hyvää OMOP-CDM mallin käytölle.

OMOP-CDM tai muut tietomallistandardit eivät ole vielä laajamittaisesti käytössä. Tähän ongelmaan sisältyy lukuisia tekijöitä. Sähköinen potilastieto on luonteeltaan heterogeenistä, minkä vuoksi potilastiedot eivät sijaitse vain yhdessä lähdejärjestelmässä. Hajallaan säilytetyn potilastiedon muuntaminen yhteensopivaksi OMOP-CDM-standardin kanssa vaatii analyytikoilta huomattavasti työtä, mikä hidastaa integraatiota ja on kallista toteuttaa.

Yhtenä ratkaisuna tähän voisi olla alkuperäisen datan tallentaminen jo valmiiksi rakenteisessa muodossa. Rakenteinen kirjaaminen tekisi esimerkiksi potilaskertomusten tekstidatasta järjestetympää, ja helpottaisi yhteensovittamista OMOP-CDM:n kanssa. Jotkin potilastietojärjestelmät ovat yrittäneet ratkaista ongelmaa laajemman rakenteisen kirjaamisen kautta. Esimerkkinä tällaisesta järjestelmästä on Suomessa hiljattain käyttöön otettu Apotti. Lääketieteen kirjaamiskäytännöt ovat pitkälti vakiintuneet nykyisenlaisiksi

vapaaksi tekstiksi rakenteisuuden sijaan, joten käytännön kirjaamistyötä tekevien kliinikoiden vastaanotto esimerkiksi Apotille ei ole ollut kovin positiivinen. Näin ollen vaihtoehdoksi jää teknologioiden ja järjestelmien kehittäminen, jotka muuttavat järjestämättömän tiedon OMOP-CDM -yhteensopivaksi. Tästä esimerkkinä on mm. Sharman ym. kehittämä fenotyypitystyökalu [17]. Koska nämä eivät toistaiseksi sovellu suoraan yleistettäväksi muiden alojen käyttöön, tarvitaan vielä runsaasti jatkokehitystä joko uusien vastaavien työkalujen tai paremmin yleistettävien versioiden saavuttamiseksi.

Motivaationa OMOP-CDM:n yleistämiseksi voidaan pitää paremman potilasdatan ja sitä kautta laadukkaamman kvantitatiivisen ja paremmin yleistettävän tutkimustiedon kerryttämistä. Lisäksi OMOP-CDM mahdollistaa tutkimustyön yhteensovittamisen paitsi laitostyö myös valtiörajojen yli, kuten Vashishtin [21] tutkimuksesta nähdään. Maininnanarvoista tutkimuksesta tekee sen kattava tietoaineisto, joka on kerätty kahdeksasta lähdejärjestelmästä kolmesta eri valtiosta. Valtaosa nykyisestä kliinisestä tutkimuksesta tapahtuu kuitenkin varsin paikallisesti tai vähintään valtiörajojen sisällä, jolloin tutkimustuloksia tulkitessa on harkittava tarkkaan, soveltuvatko saavutetut tutkimustulokset myös esimerkiksi muiden maiden väestöihin sovellettavaksi. Jos käytettävissä olisi paremmin saatavilla olevaa dataa, kuten OMOP-CDM:n on tarkoitus mahdollistaa, voitaisiin tehdä enemmän paremmin yleistettävää tutkimusta.

Johnstonin ja hänen työryhmänsä tutkimus [9] toimii erinomaisena esimerkkinä OMOP-CDM:n ja OHDSI:n työkalujen soveltamisesta lääketieteellisen datan analysoinnissa. Tutkimus projisoi OMOP-CDM:n arvoa potilashoidon päätöksenteossa osoittamalla kuinka sen avulla voidaan laajentaa tietoaineistoa ja tukea yksilöllisiä terveydenhuollon ratkaisuja. Heidän kehittämänsä patient-level prediction (PLP)-malli, joka ennustaa lihavuusleikatun tyypin 2 diabeetikon lääkehoidon tarvetta, perustuu yli 16 000 potilaan tietoaineistoon, korostaen OMOP-CDM:n kykyä yhtenäistää ja analysoida laajoja tietoaineistoja eri lähdejärjestelmistä. Tämä tutkimus ei ainoastaan paranna potilaskohtaisten hoitosuunnitelmien laatua, vaan myös edistää OMOP-CDM:n käyttöä terveydenhuollon päätöksenteon tietomallina, joka mahdollistaa tehostetun tietojen käsittelyn ja kannustaa kehittämään vastaavia työkaluja kliinikoiden tueksi.

Esteenä OMOP-CDM:n laajemmalle käyttöönotolle lienee pääasiassa tietoisuuden ja rahoituksen rajallisuus. Monet tutkijat eivät ole tietoisia OMOP-CDM:n tarjoamista hyödyistä, ja vaikka olisivatkin, heiltä saattaa puuttua tekninen osaaminen lähdedatan muuttamiseen OMOP-yhteensopivaksi. Tämän vuoksi tutkimusryhmiin täytyisi palkata data-analyttikkoja, mikä luonnollisesti vaatii lisää taloudellista resursointia. Taloudellisten hyötyjen saavuttaminen laajoista aineistoista voi kuitenkin viipyä, sillä edellytyksenä on, että tutkimustulokset todella johtaisivat hoitokäytäntöjen tehostumiseen. On myös tärkeää ylläpitää tutkimuksen eettisyyttä ja käsitellä potilastietoja anonymisti, erityisesti kansainvälisissä tutkimuksissa. Tätä OMOP-CDM on pyrkinyt ratkaisemaan sillä, että tietomalliin ei tallenneta mitään potilasta yksilöiviä tietoja [15].

7. YHTEENVETO

Tämä kandidaatintutkielma keskittyy esittelemään avoimen yhteisön kehittämää Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM):ia ja sen käyttöä lääketieteellisen datan analytiikassa ja hallinnassa. OMOP-CDM tarjoaa innovatiivisen ratkaisun potilastietojen hyödyntämiseen suljettujen systeemien ulkopuolella, sekä edistää tiedon jakamista ja käyttöä lääketieteellisessä tutkimuksessa. Tutkielma toteutettiin kirjallisuuskatsauksena tarkastelemalla julkaisuja kattavasti eri julkaisutietokannoista.

OMOP-CDM on standardoitu tietomalli, joka on suunniteltu lääketieteellisen tiedon rakenteelliseen järjestämiseen ja hallintaan. Tietomallilla tarkoitetaan yleisesti muodollista tapaa, jolla tietoalkiot ja niiden väliset suhteet esitetään. Erityisesti terveydenhuollon data-analytiikassa tietomallit ovat tärkeässä asemassa, kun sähköisiä potilastietoja pyritään organisoimaan ja hallitsemaan tehokkaasti. Tietojärjestelmien moninaisuuden vuoksi terveydenhuollossa käytettävien tietomallien määrä on suuri. OMOP-CDM erottuu joukosta tavoitteellaan standardoida potilastiedot yhdenmukaisiksi eri järjestelmien kesken. Tämä standardointi ei ainoastaan helpota tietojen integrointia ja vertailua eri lähteistä, vaan mahdollistaa laajemman, korkealaatuisen kvantitatiivisen tutkimuksen toteuttamisen, mikä on olennaista lääketieteen edistykselle.

Observational Health Data Sciences and Informatics (OHDSI) -yhteisö tukee OMOP-CDM:n käytön edistämistä kehittämällä ja ylläpitämällä avoimen lähdekoodin sovelluksia, jotka helpottavat potilastietojen muuntamista OMOP-CDM-standardin mukaiseen formaattiin. OHDSI toimii lääketieteen ja datatieteen rajapinnassa, edistäen kansainvälistä yhteistyötä potilastietojen yhtenäistämässä ja analysoinnissa. Datatieteen kasvava rooli lääketieteellisessä tutkimuksessa on avannut uusia mahdollisuuksia laajojen ja monimutkaisten tietoaineistojen hallintaan ja analyysiin. Lääketieteen ja datatieteen asiantuntijoiden välinen yhteistyö auttaa ratkaisemaan ja ymmärtämään sekä yksilöiden että populaatioiden terveysongelmia entistä paremmin.

Työssä käsiteltiin kolmea tapaustutkimusta, jotka osoittavat OMOP-CDM:n hyötyjä lääketieteellisessä tutkimuksessa. Ensimmäisessä tutkimuksessa kehitettiin potilastason ennustemalli lihavuusleikkauspotilaiden lääkehoidon tarpeen ennustamiseksi hyödyntäen OMOP-CDM:n avoimen lähdekoodin sovelluksia ja OHDSI-yhteisön työkaluja. Toisessa tutkimuksessa yhdistettiin useiden laitosten potilastiedot OMOP-CDM:n avulla ja luotiin

laajan kansainvälisen aineiston tyypin 2 diabeteksen lääkehoitovaihtoehtojen vertailuun. Kolmas tutkimus keskittyi kehittämään OMOP-CDM:ää hyödyntävän fenotyypitystyökalun. Tämän työkalun avulla oli mahdollista koota ja analysoida potilastietoja erilaisten liitännäissairauksien perusteella.

Tässä kandidaatintutkielmassa arvioitiin OMOP-CDM:n roolia ja sen haasteita terveydenhuollossa ja lääketieteellisessä data-analytiikassa. Pohdinnassa keskityttiin OMOP-CDM:n käyttöönoton esteisiin, kuten heterogeeniseen sähköiseen potilastietoon ja sen standardoinnin haasteisiin. Lisäksi käytiin läpi OMOP-CDM:n mahdollisuuksia helpottaa kliinistä päätöksentekoa ja sen potentiaalia kansainvälisissä tutkimuksissa. Lopuksi tuotiin esiin tarve lisätutkimukselle ja kehitykselle, erityisesti teknologian ja eettisten näkökohtien osalta, jotta OMOP-CDM:n soveltuvuutta voidaan laajentaa.

LÄHTEET

- [1] UK Health Data Research Alliance. *Recommendations for Data Standards in Health Data Research*. Marraskuu 2021. URL: <https://ukhealthdata.org/wp-content/uploads/2021/12/211124-White-Paper-Recommendations-of-Data-Standards-v2-1.pdf> (viitattu 24. 11. 2023).
- [2] Catherine Cote. *Applications of Data Analytics in Health Care*. 2021. URL: <https://online.hbs.edu/blog/post/data-analytics-in-healthcare> (viitattu 07. 11. 2023).
- [3] Raquel Dias ja Ali Torkamani. "Artificial intelligence in clinical and genomic diagnostics". *Genome medicine* 11.1 (2019), s. 1–12.
- [4] Martin Dugas ym. "Portal of medical data models: information infrastructure for medical research and healthcare". *Database* 2016 (2016), bav121.
- [5] EHDEN. *A federated network of Data Partners*. 2022. URL: <https://www.ehden.eu/datapartners/> (viitattu 24. 11. 2023).
- [6] Timothy C Guetterman. "Basics of statistics for primary care research". *Family medicine and community health* 7.2 (2019).
- [7] *How the Vocabulary is Built*. URL: <https://ohdsi.github.io/CommonDataModel/vocabulary.html> (viitattu 22. 11. 2023).
- [8] Md Saiful Islam ym. "A systematic review on healthcare analytics: application and theoretical perspective of data mining". Teoksessa: *Healthcare*. Vol. 6. 2. MDPI. 2018, s. 54.
- [9] Stephen S Johnston ym. "Using machine learning applied to real-world healthcare data for predictive analytics: an applied example in bariatric surgery". *Value in health* 22.5 (2019), s. 580–586.
- [10] Tarja Laitinen Arho Virkki ja Kimmo Porkka. "FinOMOP: terveystietojen kansainvälinen harmonisointi". *Duodecim* 138 (2022), s. 1761–3.
- [11] Janne Lepistö ja Timo Ukkola. *Asiakas- ja potilastietojärjestelmien tilannekuva ja sen analyysi 2020*. 2020. URL: https://www.kuntaliitto.fi/sites/default/files/media/file/APTJ-tilannekuva2020_AKUSTI110620_0.pdf (viitattu 22. 11. 2023).
- [12] Hao Liu ym. "Ontology-based categorization of clinical studies by their conditions". *Journal of Biomedical Informatics* 135 (2022), s. 104235.
- [13] *OHDSI software tools*. URL: <https://ohdsi.org/software-tools/> (viitattu 22. 11. 2023).
- [14] *OMOP CDM Background*. URL: <https://ohdsi.github.io/CommonDataModel/background> (viitattu 25. 09. 2023).
- [15] *Preserving Privacy in an OMOP CDM Implementation*. URL: <https://ohdsi.github.io/CommonDataModel/cdmPrivacy.html> (viitattu 30. 11. 2023).

- [16] Chandan K Reddy ja Charu C Aggarwal. "Healthcare data analytics". Teoksessa: vol. 36. CRC Press, 2015. Luku Chapter 1, s. 1–8.
- [17] Himanshu Sharma ym. "Developing a portable natural language processing based phenotyping system". *BMC Medical Informatics and Decision Making* 19.3 (2019), s. 79–87.
- [18] *Standardized Data: The OMOP Common Data Model*. URL: <https://www.ohdsi.org/data-standardization/> (viitattu 25. 09. 2023).
- [19] Sri Venkat Gunturi Subrahmanya ym. "The role of data science in healthcare advancements: applications, benefits, and future prospects". *Irish Journal of Medical Science (1971-)* 191.4 (2022), s. 1473–1483.
- [20] *The Book of OHDSI*. URL: <https://ohdsi.github.io/TheBookOfOhdsi> (viitattu 23. 11. 2023).
- [21] Rohit Vashisht ym. "Association of hemoglobin A1c levels with use of sulfonylureas, dipeptidyl peptidase 4 inhibitors, and thiazolidinediones in patients with type 2 diabetes treated with metformin: analysis from the observational health data sciences and informatics initiative". *JAMA network open* 1.4 (2018), e181755–e181755.
- [22] *What is a Data Model?* URL: <https://cedar.princeton.edu/understanding-data/what-data-model> (viitattu 15. 10. 2023).