Analiza danych w ramach EHDS

Między dostępnością danych a wiarygodnością wyników

R. Topór-Mądry

Interdyscyplinarne Centrum Danych o Zdrowiu UJCM

Jakie dane udostępni EHDS do użycia wtórnego? (Art. 51)

Kategorie danych, które data holders MUSZĄ udostępnić na potrzeby badań, innowacji, polityki zdrowotnej, nadzoru i AI

a) EHR

Wyniki lab., obrazowanie, karty wypisu, e-recepty

III 2029

b) Dane admin.

Roszczenia, dyspensacje, dane rozliczeniowe

III 2029

c) Rejestry

Nowotworowe, chorób rzadkich, szczepień

III 2029

d) Urządzenia med.

IoMT, wearables, aplikacje wellness

III 2029

e) System zdrowia

Potrzeby, zasoby, wydatki, dostęp

III 2029

f) Personel med.

Status, specjalizacja, instytucja

III 2029

g) Dane genomowe

Sekwencjonowanie, biobanki

III 2031

h) Dane omiczne

Proteomika, metabolo-mika, lipidomika

III 2031

i) Badania kliniczne

Wyniki trialów, kohorty, ankiety zdrowotne

III 2031

j) Determ. społeczne

Socjoekonomiczne, środowiskowe

III 2031

k) Substancje

Pozwolenia, zgłoszenia działań niepożądanych

III 2029

l) Krajowe +

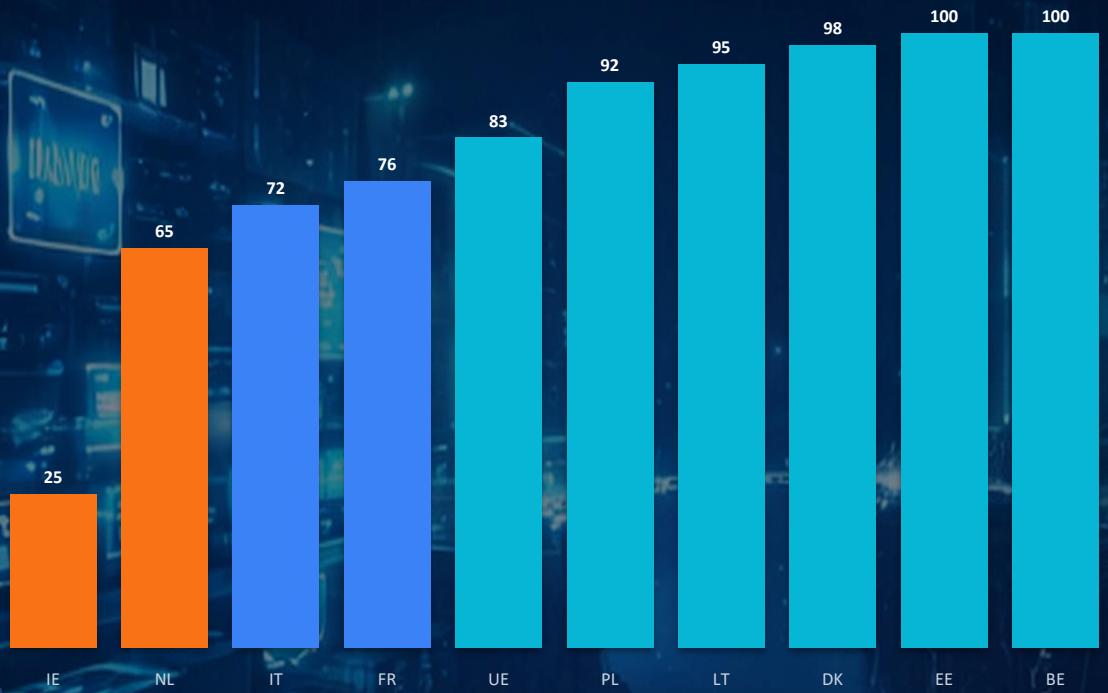
Państwa mogą rozszerzyć katalog

elastyczne

[1] Rozporządzenie (UE) 2025/327, Art. 51 i Art. 105.

Przepaść cyfrowa między krajami UE

eHealth Maturity Score 2024 – rozpiętość od 25% do 100%



Konsekwencje

- Obrazowanie med. dostępne w 26% krajów
- Prywatni świadczeniodawcy podłączeni w 59%
- Dane z tak różnych systemów będą miały fundamentalnie różną jakość
- 'Lepsze dane' zależy od celu (fitness-for-purpose)

[2] EC Digital Decade 2025: eHealth Study. [3] Euronews Feb 2026.

Wyzwania dostępności i jakości danych w EHDS

Jakość w obrębie krajów	Szpitale akademickie vs rejonowe – różne systemy EHR, kodowanie, incentywy dok. (DRG vs badania). Praktyka copy-paste propaguje błędy. Sektor prywatny podłączony w 59% [4,5].
Etykieta jakości (Art. 78)	6 elementów: dokumentacja, jakość techniczna, procesy QM, pokrycie, dostęp, modyfikacje. DOBROWOLNA dla większości (obowiązkowa tylko przy finansowaniu UE). Self-labelling – HDAB weryfikuje na żądanie, nie z urzędu [6].
Projekt QUANTUM	35 partnerów UE. Specyfikacje etykiety Art. 78 do III 2027. Model dojrzałości, pilotaż na EHR/genomice/rejestrach. Inspiracja: HDR-UK, OHDSI/EHDEN. Brak jednego standardu (EHDS nie narzuca OMOP/FHIR) [7].
Opłaty i ryzyko finansowe	Art. 62: HDAB + data holders pobierają opłaty za ocenę, przygotowanie, anonimizację, SPE. Brak mechanizmu zwrotu za dane niskiej jakości. Data permit: do 2+2 mies. Postulat: losowe próbki PRZED zobowiązaniem [8].
SPE, opt-out, discovery	DWA TRYBY DOSTĘPU: (1) Data permit (Art.68) → praca w SPE (Art.73): dane pseudo/ anonimizowane, eksport TYLKO anonimowych wyników. (2) Data request (Art.69) → HDAB sam przetwarza i dostarcza zagregowane statystyki – SPE nie jest wymagane. Opt-out (Art. 71): masowe wycofania = spadek reprezentatywności [9].

[4] OECD WP 160. [5] de Lusignan 2018. [6] Art.78 EHDS. [7] QUANTUM. [8] Art.62 EHDS. [9] Forster 2025 EJPH.

Konsekwencje danych niskiej jakości

Masowe użycie danych zdrowotnych bez kontroli jakości prowadzi do tego, że powstają:

Fałszywe zależności

Błędne kodowanie ICD → artefaktowe korelacje. Misklasyfikacja outcome zmienia estymację ryzyka względnego. Np. czułość kodu zawału serca: 75-95%, ale dla astmy: 30-55%. **Badacz nie wie, jak bardzo kody są niedokładne** [10].

Niereprezentatywne wyniki

Dane EHR odzwierciedlają korzystanie z usług, nie stan zdrowia populacji. **Osoby bezdomne, nieubezpieczone, imigranci – niedoreprezentowani.** Jeffries: tylko 13,4% badań analizowało rasę. **Opt-out pogarsza problem** [11].

Błędne algorytmy AI

Modele trenowane na danych z biasem selekcji utrwalają i WZMACNIAJĄ nierówności. Pulsoksymetry: 3× ryzyko hipoksemii u czarnoskórych (Sjoding, NEJM 2020). Obermeyer: algorytm na >100 mln pacjentów – 26% więcej chorób przy tym samym score [12].

Niereprodukowalne badania

Różne metody ekstrakcji z tego samego EHR dają różne wyniki. De Vries: te same dane GPRD → różne RANKINGI szpitali. Hripcsak: 10 ośrodków, te same kohorty – różne wyniki. **Bez standaryzacji – brak reprodukowalności** [13].

[10] Goldstein 2022. [11] Jeffries JAMIA 2023. [12] Sjoding NEJM 2020. [13] de Vries GPRD.

Wymiary jakości danych – w tym proweniencja

Wg Kahn et al. (2016), DAMA-DMBOK, Wang & Strong (1996), Art. 78 EHDS. Jakość NIE jest obiektywna – zależy od celu (fitness-for-purpose).

Kompletność
Brak danych diagnoz, procedur, wypisu

Dokładność
Błędy ręczne, copy-paste, nieaktualny stan

Spójność
Data zgonu < data przyjęcia, logiczne sprzeczności

Aktualność
Opóźnienia we wprowadzaniu, stare formularze

Unikalność
Duplikaty przy łączeniu wielu źródeł

Reprezentatywność
Opt-outy, brak mniejszości, niedostępność opieki

Walidowalność
Możliwość chart review, krzyżowa z rejestrami

Pochodzenie
Skąd dane? kto je zbierał, jakie transformacje przeszły, jak łączono zbiory. Bez proweniencji – brak reprodukowalności [18].

[18] Kahn MG et al. EGEMS 2016;4(1):1244. Art. 78(3) EHDS: proweniencja jako element etykiety.

Rola twórców danych w ocenie jakości

Zalety zaangażowania

- Wiedza kontekstowa: DLACZEGO dane są niekompletne
- Znajomość incentywów dokumentacyjnych
- Identyfikacja artefaktów (zmiana systemu EHR)
- Chart review najskuteczniejszy gdy prowadzony przez osoby znające kontekst kliniczny

VS

Ryzyka i ograniczenia

- Konflikt interesów: self-labelling (Art. 78)
- Obciążenie: personel i tak przeciążony
- Subiektywność: różni klinicysci różnie oceniają
- Rozwiązanie: NIEZALEŻNY audyt + konsultacja z twórcami jako UZUPEŁNIENIE
- Losowe próbki do chart review = STANDARD

▶ REKOMENDACJA: Niezależny audyt zewnętrzny + konsultacja z twórcami danych (nie self-labelling). Losowe próbki do chart review = STANDARD. Interdyscyplinarny zespół: klinicysta + epidemiolog + informatyk + statystyk.

[19] HealthData@EU Pilot. EJPH 2025;35(S3):iii32-38.

AI i złe dane – przypadki z praktyki

Obermeyer (Science 2019)

Algorytm >100 mln pacjentów: koszty jako proxy potrzeb zdrowotnych. Czarnoskórzy o 26% więcej chorób przy tym samym score. 17,7% vs 46,5% zakwalifikowanych. Korekta: zmiana proxy zmniejszyła bias o 84% [22].

Epic Sepsis Model (JAMA 2021)

Własnościowy model w setkach szpitali. Deklarowane AUC: 0,76–0,83. Walidacja zewnętrzna (Wong): AUC=0,63, czułość 33%. Przeoczał 67% seps, alarmy w 18% hospitalizacji [23].

Dermatologia AI

CNN trenowane na jasnej karnacji – istotnie niższa dokładność diagnozy na ciemnej skórze. 23% wyższy wskaźnik fałszywie negatywnych dla populacji wiejskich [24].

[22] Obermeyer Z. Science 2019;366:447. [23] Wong A. JAMA IM 2021;181:1065. [24] Adamson JAMA Derm 2018. [25] Vyas NEJM 2020;383:874.

Dane retrospektywne z EHDS – pozycja w hierarchii dowodów

Dlaczego same dane EHR nie wystarczają – i dlaczego RCT pozostaje złotym standardem

Dlaczego dane retrospektywne z EHR są niewystarczające do każdej analizy?

❌ Brak randomizacji – niezmierzone confoundery (DNR/DNI, styl życia, adherencja)

❌ Brak kontroli nad zbieraniem – badacz dostaje to, co ktoś kiedyś wpisał, nie to czego potrzebuje

❌ Selection bias – EHR = tylko osoby korzystające z systemu (nie cała populacja)

❌ Informational bias – czułość ICD: 30–90%

❌ Immortal time bias, prevalent user bias

❌ Brak ślepej próby – nie można ustalić przyczynowości

▶ HRT (WHI odwróciło wyniki), beta-karoten (+18% nowotworów), Xigris (wycofany po 10 latach) [1–4]

[1] Burns PB et al. J Hand Surg Am 2011;36:753. [2] Costa et al. Front Public Health 2025. [3] Rossouw JAMA 2002;288:321. [4] Ranieri NEJM 2012;366:2055.

ML vs RCT vs badanie prospektywne – porównanie

Machine learning NIE poprawia jakości danych wejściowych – „garbage in, garbage out”

ML na danych EHR (retrospektywne)	RCT (złoty standard)	Badanie prospektywne obserwacyjne
Tylko to, co ktoś wpisał. Brak kontroli kompletności.	Celowo wg protokołu. Randomizacja. Standaryzacja.	Celowo wg CRF. Bez randomizacji, pełna kontrola.
Zmienne zakłócające NIEWIDOCZNE – ML ich nie znajdzie, bo nie ma w danych.	Eliminowane randomizacją (znane i nieznane).	Mierzone i kontrolowane, ale nieznane mogą pozostać.
Przyczynowość? NIE! ML = korelacje. Może znaleźć artefakty.	TAK – jedyna metoda ustalenia przyczynowości.	Ograniczona – silniejsza niż retro, ale nie dorównuje RCT.
"Garbage in → garbage out" ML wyciąga wzorce z BŁĘDÓW.	Najwyższa – podwójne wprowadzanie, monitoring.	Wysoka – CRF dedykowane, ale bez audytu RCT.
Nature Med 2025: OS o 3 mies. niższa. 40% AI w RCT bez korzyści klinicznych [5–7].	Złoty standard, ale kosztowny, długi, ograniczona generalizowalność.	Dobry kompromis: celowe zbieranie + szersza populacja. Wymaga walidacji w RCT.

WNIOSEK: ML jest narzędziem do analizy – NIE zastępuje jakości danych. Prospektywne > retrospektywne. RCT > wszystko.

[5] NCBI Bookshelf 2023: EHR as Research Data. [6] Nature Med 2025: TrialTranslator. [7] npj Digit Med 2021: 65 RCTs z AI.

Podsumowanie

⚠ WYZWANIA

Fundamentalnie nierówna jakość

eHealth Maturity: od 25% (IE) do 100% (BE/EE). Różne systemy EHR, kodowanie, incentywy. Porównywalność fikcyjna bez standaryzacji.

Badacz ponosi ryzyko – brak zwrotu

Opłaty Art.62, data permit do 2+2 mies. Brak mechanizmu rekompensaty za dane niskiej jakości. Konieczne: losowe próbki PRZED aplikacją.

EHR nie zawiera kluczowych informacji

Przyczyny decyzji, adherencja, styl życia, kontekst społeczny – niedostępne. Dane retrospektywne = słabe dowody. AI wzmacnia biasy, nie koryguje.

✔ SZANSE

Bezprecedensowa skala danych zdrowotnych

EHDS udostępni dane z 27 krajów, setki milionów pacjentów. Część krajów (DK, EE, BE, PL) ma dojrzałe systemy EHR z danymi WYSTARCZAJĄCEJ jakości do wielu analiz.

Mechanizmy poprawy jakości w budowie

Art.78 + QUANTUM: etykieta jakości z 6 wymiarami. STaRT-RWE i RECORD: standardy raportowania. OHDSI/EHDEN: modele danych (OMOP CDM) już sprawdzone w praktyce.

Nowe możliwości badawcze

Choroby rzadkie: wystarczająca liczebność dzięki skali UE. Farmakoepidemiologia, nadzor post-marketingowy, bezpieczeństwo leków

WNIOSEK: EHDS to ogromna szansa – ale wymaga świadomości ograniczeń.
Klucz: dobieranie pytań badawczych DO jakości danych, nie odwrotnie.
Tam, gdzie dane są dobre i wystarczające – EHDS zmieni medycynę.