ARTYKUŁY POGLĄDOWE (REVIEW PAPERS)

Metody analizy danych w ochronie zdrowia - definicje, interpretacje prawne

(Data analysis methods in health care - definitions, legal interpretations)

M Machota [1,A,D], S Kasza [2,A,D], A Romaszewski [2,E], Z Kopański [2,D,F], W Uracz [1,B,E], F Furmanik [1,C], S Dyl [1,B], J Tabak [1,B]

1. Collegium Masoviense – Wyższa Szkoła Nauk o Zdrowiu
2. Wydziału Nauk o Zdrowiu Collegium Medicum Uniwersytet Jagielloński

Abstract - The authors assumed that the starting point in the considerations devoted to new methods of data analysis in health care is to define the basic concepts of the discussed topic. They focused on defining them, including legal aspects, data concepts, information concepts, knowledge, Big Data, three "V", data warehouse, Internet of Things, data processing, personal and sensitive data.

Key words - methods of data analysis, basic concepts, definitions .

Streszczenie - Autorzy wyszli z założenia, że punktem wyjścia w rozważaniach poświęconych nowym metodom analizy danych w ochronie zdrowia, jest zdefiniowanie podstawowych pojęć z zakresu omawianej tematyki. Skupili się na zdefiniowaniu , w tym uwzględnieniu aspektów prawnych, pojęcia danych, pojęcia informacji, wiedzy, Duże Dane, trzy „V", hurtownie danych, Internet Rzeczy, przetwarzania danych, danych osobowych oraz wrażliwych.

Słowa kluczowe - metody analizy danych , podstawowe pojęcia, definicje.

Wkład poszczególnych autorów w powstanie pracy— A-Koncepcja i projekt badania, B-Gromadzenie i/lub zestawianie danych, C-Analiza i interpretacja danych, D-Napisanie artykułu , E-Krytyczne zrecenzowanie artykułu, F-Ostateczne zatwierdzenie artykułu

Adres do korespondencji — Prof. dr Zbigniew Kopański, Wydziału Nauk o Zdrowiu Collegium Medicum Uniwersytet Jagielloński, Kraków, ul. Piotra Michałowskiego 12, PL-31-126 Kraków, e-mail: zkopanski@o2.pl

Zaakceptowano do druku: 29.08.2018.

DEFINICJA POJĘĆ

Punktem wyjścia w rozważaniach poświęconych nowym metodom analizy danych w ochronie zdrowia, jest zdefiniowanie podstawowych pojęć z zakresu omawianej tematyki.

Pierwszym sformułowaniem, które wymaga zdefiniowania jest pojęcie danych. Dane według ustawy o systemie informacji w ochronie zdrowia to litery, wyrazy, cyfry, teksty, liczby, znaki, symbole, obrazy, kombinacje liter, cyfr, liczb, symboli i znaków, zebrane w zbiory o określonej strukturze, dostępne według określonych kryteriów, w tym dane osobowe.

W ramach rozważań nad terminologią warto też przytoczyć definicje baz danych, która nierozerwalnie łączy się z omawianym tematem i jest często przytaczana. W polskim prawodawstwie definicja baz danych została sformułowana w ustawie o ochronie baz danych. W rozumieniu ustawy baza danych oznacza zbiór danych lub jakichkolwiek innych materiałów i elementów zgromadzonych według określonej systematyki lub metody, indywidualnie dostępnych w jakikolwiek sposób, w tym środkami elektronicznymi, wymagający istotnego, co do jakości lub ilości, nakładu inwestycyjnego w celu sporządzenia, weryfikacji lub prezentacji jego zawartości.

Kolejnym istotnym do wyjaśnienia terminem jest pojęcie informacji. Zdefiniowanie pierwszego hasła nie jest łatwe z uwagi na mnogość definicji i teorii rozpowszechnionych przez naukowców z różnych dziedzin na całym świecie. Jednak na potrzeby tej rozprawy wystarczy określenie użyte przez Papińską-

Kacperek w opracowaniu pt. „Usługi Cyfrowe", która definiuje informacje jako każdy czynnik (także abstrakcyjny), który może być wykorzystywany przez organizmy żywe lub urządzenia automatyczne do racjonalnego działania lub sterowania. Autorka wymienia także dwa sposoby pojmowania informacji: obiektywny oraz subiektywny. „*W obiektywnym (podejście matematyczne, fizyczne, ilościowe, oparte na pojęciu entropii) – informacja oznacza pewną własność fizyczną lub strukturalną obiektów. W subiektywnym zaś – informacja ma charakter względny i jest tym co umysł może przetworzyć i wykorzystać do własnych celów" [4]. Z omawianego pojęcia wywodzi się też termin społeczeństwa informacyjnego jako obecnej struktury społecznej. Fakt ten jest istotny, ponieważ określa genezę zapotrzebowania na tytułowy temat. „Społeczeństwo informacyjne określa kolejny etap w historii świat, w którym jednostki jako - konsumenci, petenci, pracownicy i twórcy – mają możliwość, chcą i potrafią zdobyć oraz wykorzystać znalezione informacje w różnych obszarach życia codziennego*"[4].

Przedstawienie obecnej struktury społecznej jako społeczeństwa informacyjnego, nie zamyka całkowicie rozważań na temat zmian zachodzących w metodach analitycznych i kierunkach badawczych. W literaturze naukowej społeczeństwo informatyczne przedstawia się jako kolejny etap rozwoju pomiędzy bardziej rozwiniętą strukturą jaką jest społeczeństwo wiedzy. Termin ten odnosi się do populacji, w której pozyskiwani, rozwijanie, lokalizowanie, zachowanie, dzielenie się oraz wykorzystywanie zgromadzonej wiedzy stanowi główny problem rozwojowy. Zdefiniowanie pojęcia również wymaga przestudiowania wielu źródeł, bowiem można wyróżnić kilka określeń na wytłumaczenie słowa wiedza. Niezaprzeczalnym jest jednak fakt, że termin ten stanowi także kolejny element ewolucji w łańcuchu dane informacje-wiedza-mądrość, a zatem definicja będzie naturalnym rozwinięciem pojęcia informacji. Wspomnianą zależność, a jednocześnie definicja wiedzy została dobrze określona „*według Ikujiro Nonaka i Hirotaka Takeuchi, gdzie informacja jest strumieniem wiadomości, podczas gdy wiedza stanowi jego wytwór, zakorzeniony w przekonaniach i oczekiwaniach odbiorcy*" [cyt. za 4]. W wyniku powyższych zapisów można rzec, że technika analizy oparta na Big Data to zatem proces, w którym przetworzone informacje zostają przekształcone w wiedzę, która jest wykorzystywana przy podejmowaniu kluczowych decyzji, często szczebla wyższego. Dokładne zdefiniowanie pojęcia wymaga jednak dokładniejszego rozwinięcia.

Big data (pol. Duże Dane) to termin, który odnosi się do identyfikacji baz danych, których analiza i zarządzanie klasycznymi metodami jest utrudniona z uwagi na ich duży rozmiar i złożoność [2]. Definicja Big Data pojawiła się prawdopodobnie po raz pierwszy w 1998 roku w przedsiębiorstwie Silicon Graphics (SGI) i został on określony przez Johna Mashey w prezentacji pt. "Big Data and the Next Wave of InfraStress" [3]. Analiza Dużych Danych bazuje na najnowszych metodach przetwarzania: obliczenia w chmurze (ang. cloud computing), samouczenie się maszyn (ang. maschine learning), data mining, eksploracji tekstu czy wysoce zaawansowanej statystyce.

TRZY„V"

Do dokładnego określenia czym są Duże Dane zwykło się przyjmować definicję Gartnera pod postacią tak zwanych trzech „V". Z języka angielskiego są to [4-6]:

- Volume (objętość),
- Velocity (dynamika) oraz
- Variety (zróżnicowanie).

Objętość. Obecnie liczba danych znacznie przekracza możliwość ich magazynowania na pojedynczych serwerach. Rozrost wielkości wszelakich pojęć np. z obrębu medycyny jak elektroniczne rekordy medyczne, obrazy radiologiczne, kod genetyczny człowieka czy obrazowanie 3D i innych, napędzają potencjalny wzrost, z którym należy sobie poradzić. [4,7]

Dynamika. Informacje są gromadzone w czasie rzeczywistym z ogromną prędkością. Ciągły napływ nowych danych w niespotykanych do tej pory zakresach, prowadzi do powstania nowych problemów do rozwiązania. Pod pojęciem dynamiki w ochronie zdrowia można rozumieć np. regularny monitoring stanu zdrowia np. poziomu glukozy we krwi u diabetyków, ciśnienia krwi lub EKG u pacjentów z urządzeniem pomiarowym [4,5,6]

Zróżnicowanie. O zróżnicowanych danych mówi się kiedy mamy do czynienia z różnym formatem i strukturą gromadzonych informacji. Różne rekordy o charakterze strukturalnym, niestrukturalnym, półstrukturalnym oraz multimedia sprawiają, że przetwarzanie

staje się coraz bardziej skomplikowane i wymagające. [4-7]
Niektórzy naukowcy i badacze postulują jednak poszerzenie tego zbioru o kolejne pojęcie Veracity, czyli wiarygodność. Oznacza to, że wyniki i analizy są wiarygodne i wolne od wszelkich błędów, które mogłyby zaszkodzić ostatecznemu rezultatowi. Pozycja ta jest szczególnie istotna w ochronie zdrowia, gdzie często na podstawie wydobytych informacji podejmuje się ostateczne decyzje o charakterze życia lub śmierci.[4,6-8]

Tabela 1. Charakterystyka Big Data według 4V
[Opracowanie własne na podstawie 4-8]

Lp.	Pojęcie	Tłum. angielskie	Opis
1.	Obojętność	Volume	Duża liczba oraz wielkość gromadzonych danych
2.	Dynamika	Velocity	Pobieranie danych w czasie rzeczywistym
3.	Zróżnicowanie	Variety	Charakter strukturalny danych i multimedia
4.	Wiarygodność	Veracity	Dane są wolne od wad

Naszym zdaniem należy również rozwinąć niezbędne pojęcia nierozłącznie związane z Big Data, których przybliżenie pozwoli na głębsze zaznajomienie się z omawianą tematyką, są to min.: samouczenie się maszyn, ekploracja dantych, eksploracja tekstu, hurtownia danych, Internet Rzeczy oraz obliczenia w chmurze. Dokładniejsze omówienie ostatniego wątku nastąpi w dalszej części pracy, bowiem stanowi on najbardziej istotny punkt rozwoju polskiego systemu informacji medycznej.
Samouczenie się maszyn (ang. maschine learning) to dział wywodzący się z nauk zajmujących się sztuczną inteligencją (SI). Głównym zadaniem teorii jest wykorzystanie wiedzy z zakresu SI do opracowania samodzielnego systemu, opierającego się o analizę zgromadzonej wiedzy z doświadczeń (danych) uzyskanych na drodze swojej aktywności. Zastosowanie tej metodologii jest szerokie i obejmuje działy ekonomii, chemii czy medycyny, w których liczba przetwarzanych danych wykracza poza możliwości klasycznych systemów [3,9].
Pojęcie eksploracji danych (ang. data mining) może być zdefiniowane jako proces pozyskiwania wiedzy bądź informacji z dużego zasobu danych [10]. Data mining jest jedną z możliwości otrzymywania wiedzy z rozległych baz danych. Celem eksploracji jest przekształcenie danych faktycznych, tekstowych czy numerycznych w informacje. W procesie wykorzystuje się pojedyncze lub kombinowane algorytmy, które co istotne muszą same rozpoznać przetwarzane struktury danych.
Eksploracja tekstu polega na maszynowym przeobrażeniu nieustrukturyzowanego tekstu, pozyskania z niego znaczeń i przekonwertowaniu na informację ustrukturyzowaną. Następnie wyniki są analizowane bardziej tradycyjnymi metodami. Większość metod eksploracji tekstu wywodzi się z metod przetwarzania języka naturalnego [4-6,10,11].
Hurtownie danych (HD) (ang. data warehouse) są „*złożonymi systemami informatycznymi, które przetwarzają i łączą dane pochodzące z różnych źródeł w unifikowane struktury, aby nadać im jakość i formę niezbędną dla celów analitycznych*" [13]. Z uwagi na tak określoną definicję można zauważyć, że HD są abstrakcyjną złożoną strukturą, która może być narzędziem wykorzystywanym przez stanowiska kierownicze lub analityków do podejmowania określonych decyzji [6-8,11].
Internet Rzeczy (ang. Internet of Things) jest całkowicie nową koncepcją nieposiadającą jeszcze ściśle zdefiniowanego znaczenia. Opiera się ona na integracji wirtualnego świata informacji z obiektami świata rzeczywistego poprzez dołączenie do interentu nie tylko komputerów, ale także innych urządzeń lub obiektów. Przykładem mogą być urządzenia RFID (Radio Frequency Identification), urządzenia kuchenne czy akcesoria odzieżowe (ang. wearables) [4].
Ostatnimi elementami poświęconymi definicjom są pojęcia przetwarzania danych, danych osobowych oraz wrażliwych. Według ustawy o ochronie danych osobowych, przez przetwarzanie danych rozumie się jakiekolwiek operacje wykonywane nadanych osobowych, takie jak zbieranie, utrwalanie, przechowywanie, opracowywanie, zmienianie, udostępnianie i usuwanie, a zwłaszcza te, które wykonuje się w systemach informatycznych [12]. Ta sama ustawa definiuje dane osobowe w art. 1 i dane szczególnie chronione, co do których odnosi się artykuł 27 pkt. 1 tego aktu prawnego, których przetwarzanie jest regulowane przez pkt. 2 art. 27.
„Art. 6. 1. W rozumieniu ustawy za dane osobowe uważa się wszelkie informacje dotyczące zidentyfikowanej lub możliwej do zidentyfikowania osoby fizycznej.

Osobą możliwą do zidentyfikowania jest osoba, której tożsamość można określić bezpośrednio lub pośrednio, w szczególności przez powołanie się na numer identyfikacyjny albo jeden lub kilka specyficznych czynników określających jej cechy fizyczne, fizjologiczne, umysłowe, ekonomiczne, kulturowe lub społeczne.

Informacji nie uważa się za umożliwiającą określenie tożsamości osoby, jeżeli wymagałoby to nadmiernych kosztów, czasu lub działań [12]"

„Art. 27. 1. Zabrania się przetwarzania danych ujawniających pochodzenie rasowe lub etniczne, poglądy polityczne, przekonania religijne lub filozoficzne, przynależność wyznaniową, partyjną lub związkową, jak również danych o stanie zdrowia, kodzie genetycznym, nałogach lub życiu seksualnym oraz danych dotyczących wskazań, orzeczeń o ukaraniu i mandatów karnych, a także innych orzeczeń wydanych w postępowaniu sądowym lub administracyjnym [12]."

PIŚMIENNICTWO

1. Papińska-Kacperek J. Usługi cyfrowe. Perspektywy wdrożenia i akceptacji cyfrowych usług administracji publicznej w Polsce. Łódź; Wyd. Uniwersytetu Łódzkiego, 2013.
2. Fan W, Bifet A. Mining Big Data: Current Status, and Forecast to the Future. New York; Wyd. ACM SIGKDD Explorations Newsletter 2012.
3. Diebold F. On the Origin(s) and Development of the Term "Big Data". Pier working paper archive. Pennsylvania; Penn Institute for Economic Research, Department of Economics. University of Pennsylvania, 2012.
4. Mach-Król M. Analiza i strategia Big Data w organizacjach . Stud Mater Pol Stow Zarz Wiedzą 2015; 74: 43–55.
5. Bollier D. The promise and Peril of Big Data. Raport of the Aspen Institute. Washington; Communications and Society Program, 2010.
6. Marconi K, Dobra M, Thompson C. The use of Big Data in Healthcare. In: Liebowitz J. Big Data and Business Analitics. Boca Raton; CRC Press, 2013: 229–248.
7. Schmarzo B. Big Data. Understandinga How Data Powers Big Business. Indianapolis; John Wiley &Sons Inc., 2013.
8. Chen H, Chiang R H, Storey V C. Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly 2012; 36, 4:1165–1188.
9. Cichosz P. Systemy uczące się. Warszawa; WNT, 2000.
10. Strycharz L. Eksploracja tekstu i danych – bariery prawne w Europie i Polsce. 18 luty 2016. [cytowany 8 sierpnia 2016]. Adres:http://centrumcyfrowe.pl/eksploracja-tekstu-i-danych-barieryprawne-w-europie-i-polsce/
11. Fracchia JA, Motta J, Miller LS, Armenakas NA, Schumann GB, Greenberg RA. Evaluation of asymptomatic microhematuria. Urology 1995; 46: 484–489.
12. Cormay. Twoje laboratourium. Dostęp 10.06.2017http://www.pzcormay.pl/userfiles/file/Biuletyny/TL_wiosna_2015_PL.pdf .
13. Guder W, Narayan S, Wisser H, Zawta B. Próbki: od pacjenta do laboratorium. Wpływ zmienności przedanalitycznej na jakość wyników badań laboratoryjnych. Wrocław; MedPharma Polska, 2009.
14. Alberts B. Podstawy biologii komórki. Warszawa; PWN, 2005;
15. Myśliwiec M. Choroby nerek. Warszawa; PZWL, 2008.
16. Borkowski A. Urologia. Warszawa; PZWL, 2006.