ARTYKUŁ POGLĄDOWY/REWIEV PAPER
Otrzymano/Submitted: 08.10.2016 • Zaakceptowano/Accepted: 11.12.2016

Wyzwanie współczesnego farmaceuty. „*Big Data*" – nowe spojrzenie na „stare" dane

Challenge for contemporary pharmacists. "Big Data" – a new look on "old data"

Mikołaj Mizera, Maciej Ostrowicz, Judyta Cielecka-Piontek
Katedra i Zakład Chemii Farmaceutycznej, Uniwersytet Medyczny im. Karola Marcinkowskiego w Poznaniu

Streszczenie

Big Data daje możliwość wykorzystania „starych danych" celem analizowania zmiennych zależnych i pozornie niezależnych. W pracy zaprezentowano główne wytyczne istotne dla zdefiniowania *Big Data* oraz możliwości zastosowania tego zjawiska dla rozwoju badań nad lekiem i poprawy skuteczności procedur leczenia określonych jednostek chorobowych. Autorzy szczególnie skupili się na zastosowaniu *Big Data* w pracy farmaceutów w laboratoriach naukowo-badawczych przy opracowaniu innowacyjnych molekuł oraz analizie danych otrzymywanych także podczas pracy farmaceutów w aptekach ogólnodostępnych. *(Farm Współ 2016; 9: 176-182)*

Słowa kluczowe: wielkie zbiory danych, badania nad lekiem, zdrowie publiczne

Summary

Big Data make it possible to use the "old data" in order to analyse dependent and independent variables contained in it. The paper presents the main aspects of Big Data definition as well as its possible applications in drug discovery and improvement of treatment efficacy of particular disease entities. The authors particularly focused on the use of Big Data in work of pharmacists' job in research and development laboratories (synthesis of innovative molecules) and analysis of data obtained as the result of activity of public pharmacies. *(Farm Współ 2016; 9: 176-182)*

Keywords: Big Data, studies of drugs, public health

Wstęp

Możliwość leczenia coraz większej liczby chorób jest efektem pracy w wielu obszarach badań *nauk o życiu*. Długoletnia praca naukowców, klinicystów i farmaceutów na każdym z etapów badań skupiających się na poszukiwaniu nowych sposobów leczenia lub optymalizacji przebiegu farmakoterapii określonych chorób przyczynia się do powstawania olbrzymich zbiorów danych (ang. *Big Data*). Należy podkreślić, że proces wykorzystania rozwiązań informatycznych w obszarach badań farmaceutycznych jest procesem tak dynamicznym, jak w każdej innej dziedzinie nauki i przemysłu. Trend informatyzacji nie ominął bowiem nauk farmaceutycznych, w których odpowiednie oprogramowanie odgrywa często kluczową rolę w przybliżeniu do otrzymania rozwiązania problemu badawczego. Pierwszy efektywny synergizm zastosowania rozwiązań informatycznych w badaniach nad lekiem oraz zwiększaniu jego dostępności dla określonych grup pacjentów można datować na wczesne lata osiemdziesiąte XX wieku. Już wtedy zastosowanie rozwiązań informatycznych pozwoliło zmniejszyć nakłady pracy ludzi, zwiększyć jej wydajność czy wyeliminować błędy ludzkie. Kolejne aplikacje wykorzystujące rozwinięcia informatyczne zaczęły pojawiać się w aptekach, fabrykach, szpitalach oraz ośrodkach naukowo-badawczych. Rozwój Internetu dodatkowo przyczynił się do możliwości transferowania otrzymanych wyników, co w szybkim tempie pozwoliło na stworzenie olbrzymich baz danych dla kolejno rozpatrywanych zagadnień.

Możliwość zarządzania danymi poprzez definiowanie relacji pomiędzy nimi pozwala budować zależności pomiędzy określonymi grupami zmiennych zależnych i niezależnych dla każdego obszaru badań z zakresu otrzymania i modyfikacji wybranej farmakoterapii. Gdybyśmy prześledzili ścieżkę rozwoju farmakoterapii określonej jednostki chorobowej, na każdym jej etapie otrzymano olbrzymie zbiory danych, z których tylko kilka wyników spełnia bezpośrednio zadane kryteria. Jednakże, wiedza uzyskana na każdym etapie pracy nad lekiem i badań nad efektami jego działania może zawierać dodatkowe informacje ukryte w pozornie nieskorelowanych zmiennych, których użyteczność warunkuje zastosowanie *Big Data*.

Historia *Big Data* jako dziedziny informatyki sięga lat 90., jednakże aktualnie najszerzej rozpowszechniona definicja pochodzi z raportu grupy META wydanego w 2001 roku [1]. W tymże raporcie, *Big Data* definiowana jest w oparciu o trzy filary symbolicznie oznaczone literami V (ang. *Volume, Variety, Velocity* – Objętość, Różnorodność, Szybkość). Definicja została rozszerzona w 2012 roku o konieczność wykorzystywania nowych metod analizy danych w celu zwiększenia użyteczności *Big Data* w obszarze wspomagania decyzji czy optymalizacji procesów [2]. Na rycinie 1 zaprezentowano graficzną prezentację przytoczonej definicji.

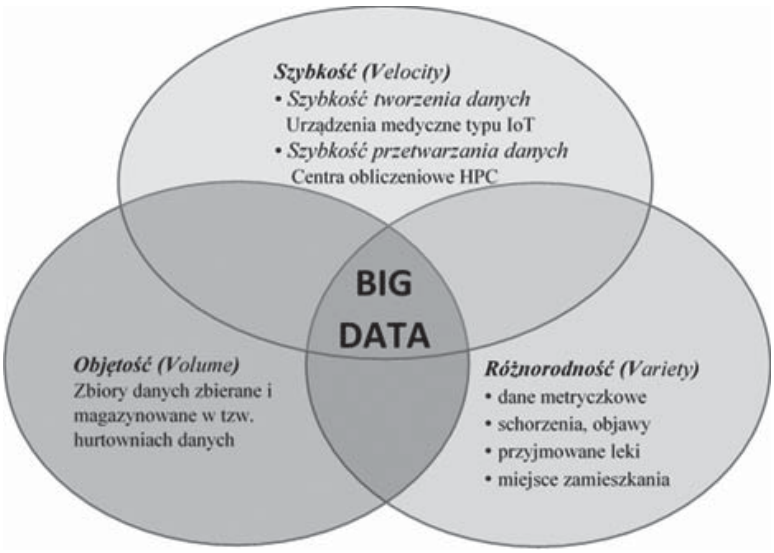
Najważniejszą własnością systemów *Big Data* jest objętość przetwarzanych danych. Jak wspomniano, każdego dnia generowane są gigabajty danych dotyczących wielu obszarów działalności farmaceutów. Dane te powstają zarówno w laboratoriach naukowo-badawczych, placówkach ochrony zdrowia jak i generowane są przez samych pacjentów używających telefonów wyposażonych w czujniki pozwalające mierzyć codzienną aktywność fizyczną czy zapisywać historię samodzielnych pomiarów np. ciśnienia, stężenia glukozy we krwi. Integracja danych z różnych źródeł (laboratoria naukowo-badawcze, szpitale, przychodnie, apteki, dane własne) dotyczących różnych aspektów badań nad lekiem i efektami jego działania wiąże się z drugim filarem *Big Data* – różnorodnością danych. Integrowane dane w ramach systemów przetwarzania mogą posiadać zarówno łatwo interpretowalny format liczbowy jak np. bezwzględne ilości substancji aktywnych farmakologicznie przyjmowanych przez pacjenta, ale także mogą przybierać trudne do klasyfikacji ontologie jak w przypadku próby opisania samego pacjenta jako punktu w zbiorze danych czy określenia substancji aktywnej w kontekście jej aktywności biologicznej. Kluczowe znaczenie dla istnienia systemu typu *Big Data* ma infrastruktura komputerów wysokiej mocy (HPC – ang. *High Power Computing*) oraz wyspecjalizowane oprogramowanie zdolne zapewnić szybkość przetwarzania danych. Pojawienie się w ostatnich latach rozwiązań informatycznych umożliwiających przeprowadzanie analizy na wielkich zbiorach danych niemalże w czasie rzeczywistym znacznie zwiększyło dostępność i użyteczność *Big Data*. Szybkość w kontekście prezentowanej definicji dotyczy również powstawania nowych danych. Do intensywnego powiększania się zbiorów danych przyczyniły się urządzenia Internetu Rzeczy (ang. IoT – *Internet of Things*) [3]. W kontekście urządzeń o przydatności medycznej, są to niewielkich rozmiarów czujniki zdolne przesyłać w czasie rzeczywistym strumień danych dotyczących parametrów życiowych pacjenta, jak np. aktualna szybkość pracy serca, poziom glukozy we krwi, czy wyniki pomiarów ze zminiaturyzowanych aparatów do pomiarów metodą Holtera. Natomiast w odniesieniu do efektów pracy laboratoriów naukowo-badawczych działających w ramach struktur globalnych, to zintegrowane systemy archiwizacji danych pierwotnych oraz ich zasoby, także po pierwotnej obróbce. Zgodnie z Wullianallur do definicji *Big data* postulowane jest dodanie czwartego czynnika „V" (*Verasity* – Wiarygodność) [4]. Wiarygodność danych odnosi się do jakości konstruowanych zbiorów danych – obecności błędów wynikających np. z niedokładności pomiarowej, niekompletności danych czy pomyłek w ich wprowadzaniu. W przypadku danych laboratoryjnych ważne jest przestrzeganie procedur pracy z wykorzystaniem określonych technik laboratoryjnych oraz praca na aparaturze pomiarowej poddanej kwalifikacji. W przypadku danych uzyskiwanych zwrotnie od pacjenta ważnym jest, aby ocenić wiarygodność danych pozyskanych samodzielne w odniesieniu do danych otrzymanych w wyniku pomiaru przez wykwalifikowany personel medyczny. W przypadkach wyodrębniania danych pozyskanych z aptek ogólnodostępnych czy szpitalnych należy uwzględniać też charakterystykę tych placówek w odniesieniu do specyfikacji położenia oraz faktu współpracy z szpitalami o zdefiniowanych profilach aktywności medycznej.

Począwszy zatem od syntezy API (ang. *active pharmaceutical ingredient*), gdzie znalezienie kompromisu pomiędzy efektywnością działania farmakologicznego *in vitro* nowych molekuł, ich toksycznością a osiągnię-

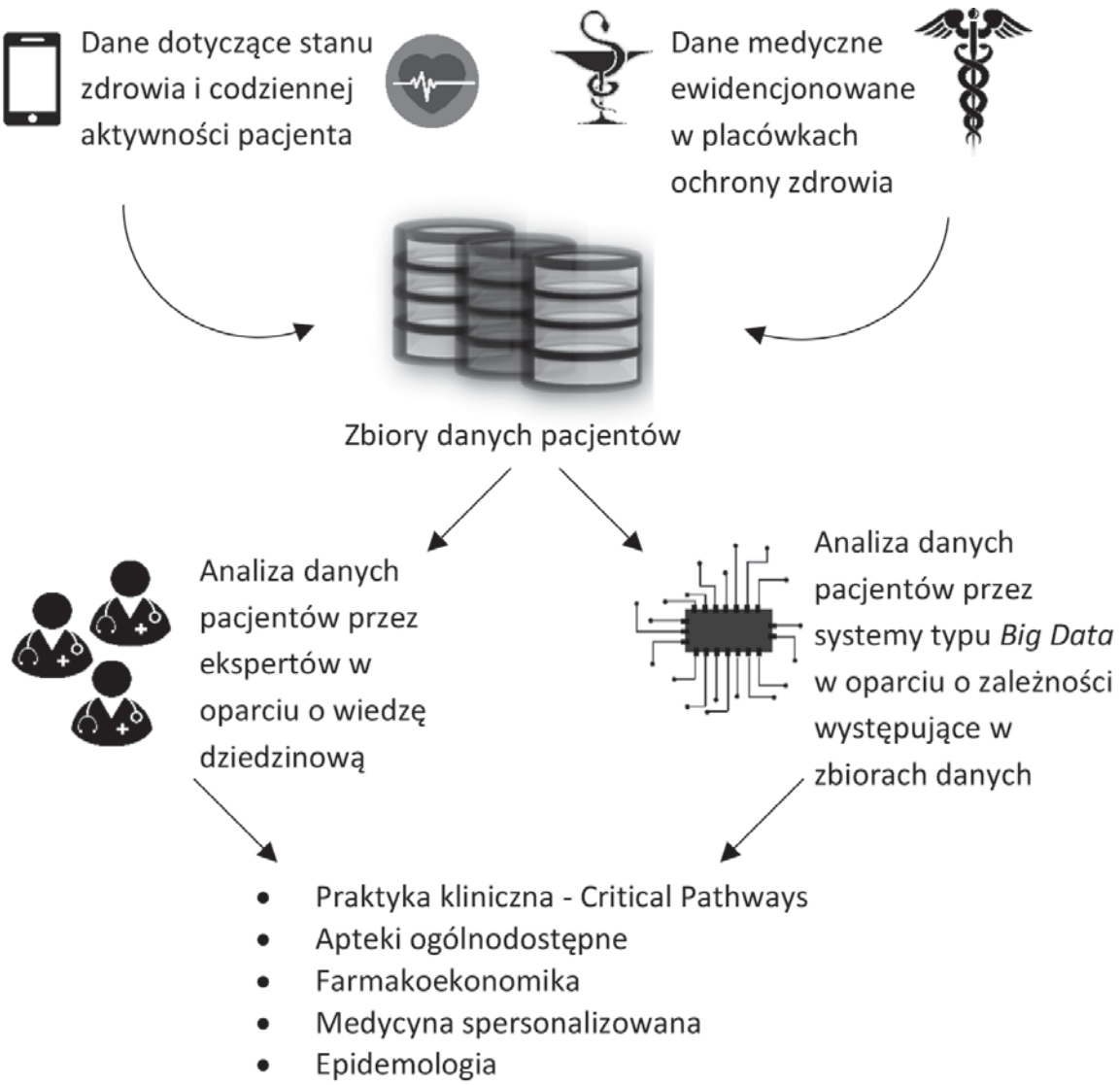
ciem określonych parametrów fizykochemicznych stanowi o sukcesie ich wprowadzenia do leczenia, a kończąc na analizie danych medycznych charakteryzujących profil danego pacjenta, pokłada się coraz większe nadzieje w wykorzystaniu „starych danych" [5]. Zastosowanie analiz Big Data, dające możliwość wykorzystania istniejących już danych, wpisuje się również w aktualne podejście dyrektywy REACH, wprowadzonej przez Unię Europejską, z początkiem czerwca 2007 roku. Rozporządzenie Parlamentu Europejskiego i Rady nr 1907/2006, zwane dyrektywą REACH (*Registration, Evaluation and Authorisation of Chemicals*) zostało przyjęte w celu zwiększenia ochrony środowiska i zdrowia przed wpływem zagrożeń, jakie mogą stanowić substancje chemiczne (w tym leki) [6]. Rozporządzenie promując nowe metody oceny zagrożeń, sprzyja zmniejszeniu liczby badań przeprowadzanych na zwierzętach oraz ustanawia procedury gromadzenia i oceny informacji na temat danej substancji i zagrożeń, jakie może ona wywoływać.

Big Data w badaniach nad nowymi lekami

Jak wspomniano, przetwarzanie wielkich zbiorów danych znalazło swoje zastosowanie także podczas prac w ośrodkach naukowo-badawczych o różnych profilach. Analiza *Big Data* jest wykorzystywana w procesie projektowania syntezy innowacyjnych cząstek w kontekście oceny ich powinowactwa do wybranych układów receptorowych. Wyzwaniem dla rozwoju badań nad nowymi API jest odkrywanie nowych cząsteczek wiodących (ang. *lead compound*) – substancji o potencjalnym zastosowaniu farmakologicznym [7]. W podejściu wspomaganym komputerowo (ang. *Computer Aided Drug Design*) stosowane są dwie główne techniki obliczeniowe. W pierwszej technice wykorzystywana jest informacja o punkcie uchwytu (ang. *structure based*), na podstawie której przeszukiwane są bazy danych ligandów i wybierane te o największym teoretycznym powinowactwie do receptora [8-9]. W drugiej z technik, wykorzystującej informację o znanych ligandach (ang. *ligand based*), definiowana jest ilościowa zależności miedzy strukturą chemiczną cząsteczki i jej aktywnością względem receptora (QSAR – ang. *Quantative Structure – Activity Relationship*) [10-12]. Zależności te, wykorzystując opisane liczbowo właściwości strukturalne cząsteczek (deskryptory) prezentują właściwości chemiczne przy pomocy modeli matematycznych. Potencjalnym zastosowaniem *Big Data* w dziedzinie poszukiwania nowych leków jest wykorzystanie informacji generowanych podczas modelowania kwantowo-chemicznego z użyciem metod przybliżających rozwiązanie funkcji falowej w kontekście tworzenia nowych, prawdopo-



Rycina 1. Trzy składowe Big Data: szybkość, objętość i różnorodność danych
Figure 1. Three components of Big Data: velocity, volume and variety



Rycina 2. Porównanie systemów wspomagania decyzji w oparciu o wiedzę ekspertów oraz w oparciu o podejście sterowane danymi

Figure 2. Comparison of decision support systems based on the knowledge of experts and based on the data-driven approach

dobnie niezrozumiałych dla badacza deskryptorów związków chemicznych. Wykorzystanie metod głębokiego uczenia maszynowego daje perspektywiczny obraz projektowania leków, gdzie największy udział w procesie badawczym będą miały dane wygenerowane przez systemy komputerowe [13]. Już teraz wykorzystywanie przez okres 20 lat zaawansowanych algorytmów przewidywania powinowactwa cząstek do receptorów oraz oceny skuteczności różnych API przyczyniło się wprowadzenia do lecznictwa takich leków jak norfloksacyna, losartan, donepezil, oseltamiwir, dorzolamid, kaptopryl, saquinavir, ritonavir, indinavir, tirofiban [14-18].

Warto też zwrócić uwagę, na możliwości, jakie dają proste, bezpieczne (np. zastosowanie testu Amesa w badaniach mutagenności) i tanie (np. zastosowanie

układu PAMPA w badaniach przenikalności) narzędzia diagnostyczne. Sprzyjają one otrzymywaniu dużej liczby danych, które po odpowiednim przetworzeniu z wykorzystaniem *Big Data* wspomagają decyzję badaczy o wyborze molekuły najbardziej bezpiecznej, ale zarazem efektywnej w odniesieniu do działania w warunkach *in vivo* [19-20].

W dobie integracji naukowej, prowadzone są także badania nad wspólnym przetworzeniem informacji wygenerowanych i publikowanych przez ośrodki badawcze na całym świecie [21]. Przykładem jest rozwijany przez fundację system OpenPHATCS mający za zadanie utworzenie i utrzymywanie otwartej platformy wymiany danych dla badań farmakologicznych. Główną ideą stojącą za powstaniem systemu jest wykorzystanie tzw. podejścia sterowanego danymi (ang. *data-driven*) w celu wskazania zależności w złożonych procesach biochemicznych warunkujących działanie farmakologiczne potencjalnych API. Baza danych tego typu umożliwia gromadzenie i analizę heterogenicznych, wielodziedzinowych danych, które w perspektywie mogą znaleźć swoje zastosowanie w odkrywaniu nowych API czy w lepszym zrozumieniu mechanizmów już zdefiniowanych jako odpowiedzialne za działanie API.

Big Data w sektorze zdrowia publicznego

W nawiązaniu do przedstawionej definicji, wykorzystanie *Big Data* w medycynie obejmuje możliwie szybkie przetwarzanie wielkich zbiorów danych, zarówno aktualnych jak i historycznych w celu wspomagania prewencji i terapii chorób pacjentów. Wspomaganie komputerowe w realizacji powyższych celów postulowane było już w latach 50, wraz z pojawieniem się pierwszych komputerów [22]. Systemy wspomagania decyzji wykorzystywały zaprogramowane reguły, które w oparciu o dane uzyskane z wywiadu lekarskiego lub automatycznej analizy wyników badań diagnostycznych prowadziły do klasyfikacji obserwacji do konkretnej jednostki chorobowej [23]. Aktualna rewolucja z wykorzystaniem *Big Data* znacząco poszerza rolę informatyki w realizacji wspomnianych celów. Na rycinie 2 przedstawiono porównanie procesu tworzenia zasad i reguł będących podstawą dla postępowań klinicznych w oparciu o historyczne wyniki badań.

W tradycyjnym podejściu, wiedza dziedzinowa reprezentowana przez lekarzy-specjalistów stanowi fundament analizy danych pacjentów, na jej podstawie przeprowadzana jest dedukcja prowadząca do postawienia diagnozy. Automatyzacja tego procesu jest możliwa poprzez zapisanie dostępnej bazy wiedzy w formie reguł logicznych tworząc system ekspercki wspomagania diagnozy (CAD – ang. *Computer Aided Diagnosis*).

Z drugiej strony, podejście sterowane danymi zakłada, iż wiedza dziedzinowa została wywiedziona na podstawie danych medycznych będących obrazem stanu zdrowia, aktywności fizycznej pacjenta a także innych, pozornie nieskorelowanych zmiennych. W takim podejściu, postawienie diagnozy opiera się na poszukiwaniu zależności między wynikami badań, obserwacjami objawów, historii spożywanego pokarmu, zażywanych leków i wielu innych niezwiązanych z postawioną diagnozą zmiennych. Każdy z tych czynników może być zapisywany w zbiorach danych pacjentów i wykorzystywany do przewidywania, ale także śledzenia stanu zdrowia i wyników prowadzonej terapii. Przykładem wykorzystania przetwarzania typu *Big Data* jest projektowanie wytycznych postępowań medycznych, takich jak ścieżki krytyczne (ang. *Critical Pathways*), rozszerzające zasady definiujące *Evidence Based Medicine* o elementy zarządzania terapią i jej optymalizacji po wdrożeniu [24]. CP wspierane przez *Big Data* może zatem obejmować monitorowanie *compliance* na poziomie apteki ogólnodostępnej, skuteczności terapii na podstawie cyklicznych spotkań z lekarzem czy stosowania się do zaleceń na podstawie monitoringu codziennej aktywności fizycznej [25]. Ogrom danych koniecznych do przetworzenia dla każdego indywidualnego pacjenta wyklucza analizę przez zespół ekspertów. Co więcej, analiza z wykorzystaniem systemów komputerów wykazuje większą skuteczność niż analiza tradycyjna. Dla przykładu, w badaniach nad wykorzystaniem antybiotyków w warunkach klinicznych analiza danych pacjentów pozwoliła na zwiększenie efektywności terapii antybiotykami poprzez identyfikację stosowania antybiotyków nieodpowiednich dla danego zakażenia lub wskazanie możliwości stosowania antybiotyków mniej kosztownych [26].

Podsumowanie

Możliwości zastosowania *Big Data* są istotne w każdym z obszarów badań nad lekiem, jak również definiowaniu efektów farmakoterapii z wykorzystaniem określonych schematów leczenia. Wykorzystanie wyników *Big Data* przez farmaceutów jest procesem, który dynamicznie się rozwija. Intensywność tego procesu jest różna dla rozpatrywanych obszarów

badań oraz skorelowana bezpośrednio z potrzebą analizy zbiorów danych o różnej wielkości i różnorodności. Nowoczesne narzędzia informatyczne zarówno programowe jak i sprzętowe są czynnikami umożliwiającymi sprawne korzystanie z zalet analizy wielkich zbiorów danych. Rozwojowi tego trendu szczególnie sprzyja cyfryzacja procesów badawczych i otrzymywania danych medycznych. Dynamiczny rozwój wspomnianych technologii pozwala sądzić, że gromadzone zbiory danych przestaną pełnić tylko funkcje retrospektywną i bazodanową. *Big Data* poszerza zakres użyteczności tych danych o aspekt predykcji, co sprawia, że dane zaczynają być rozumiane jako źródło wiedzy o zjawiskach, dzięki którym powstały.

Podziękowania/Acknowledgments
Praca naukowa finansowana ze środków Ministerstwa Nauki i Szkolnictwa Wyższego.
Nr projektu: MNISW/2016/DIR 219/NN

NAJLEPSI Z NAJLEPSZYCH

Fundusze Europejskie

Unia Europejska
Europejski Fundusz Społeczny

Praca została wykonana z wykorzystaniem Infrastruktury PL-Grid.

Konflikt interesów
Brak/None

Adres do korespondencji:
Judyta Cielecka-Piontek
Katedra i Zakład Chemii Farmaceutycznej
Uniwersytet Medyczny im. Karola Marcinkowskiego w Poznaniu,
ul. Grunwaldzka 6; 60-780 Poznań
(+48 61) 854-66-49
jpiontek@ump.edu.pl

Piśmiennictwo

1. Doug L. 3D data management: Controlling data volume, velocity and variety. META Group Research Note. 2001;6:70.
2. Beyer MA, Douglas L. The importance of 'big data': a definition. Stamford, CT: Gartner 2012;2014-18.
3. Wen-Tsai S, Chiang Y. Improved particle swarm optimization algorithm for android medical care IOT using modified parameters. J Med Syst. 2012;36(6):3755-63.
4. Wullianallur R, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst. 2014;2(1):1.
5. Ekins S. Computer Application in Pharmaceutical Research and Development. Wiley Interscience: A John Wiley and Sons, INC., Publication 2006;3(37).
6. Reach. Rozporządzenie Parlamentu Europejskiego i Rady nr 1907/2006.
7. Jadhav S, Nikam K, Gandhi A i wsp. Applications of computer science in Pharmacy: An overview. Nat J Physiol Pharm Pharmac. 2012;2(1):1-9.
8. Bin Ch, Butte AJ. Leveraging big data to transform target selection and drug discovery. Clin Pharmacol Ther. 2016; (3):285-97.
9. Amzel M. Structure-based drug design. Curr Opin Biotech. 9(4):366-9.
10. Cherkasov A, Muratov E, Fourches D i wsp. QSAR modeling: where have you been? Where are you going to? J Med Chem. 2014;57(12): 4977-5010.
11. Śliwoski G. Computational methods in drug discovery. Pharmacol Rev. 2014;66(1):334-95.
12. Bandameedi R. Provenance of Computers in Pharmacy. Clin Pharmacol Biopharm. 2016;5:153.
13. Izhar W, Dzamba M, Heifets A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. arXiv preprint 2015;arXiv:1510.02855.
14. Koga H, Itoh A, Murayama S i wsp. Structure-activity relationships of antibacterial 6, 7-and 7, 8-disubstituted 1-alkyl-1, 4-dihydro-4-oxoquinoline-3-carboxylic acids. J Med Chem. 1980;23(12):1358-63.
15. Gaurab B. How the antihypertensive losartan was discovered. Expert opinion on drug discovery 2006;1(6):609-18.

16. Lew W, Chen X, Kim C. Discovery and development of GS 4104 (oseltamivir): an orally active influenza neuraminidase inhibitor. Curr Med Chem. 2016;7(6):663-72.
17. Sugimoto H, Yamanishi Y, Ogura H i wsp. Discovery and development of donepezil hydrochloride for the treatment of Alzheimer's disease. Yakugaku Zasshi. 1999;119(2):101-13.
18. Sliwoski G, Kothiwale E, Meiler J i wsp. Computational methods in drug discovery. Pharmacol Rev. 2014;66(1):334-95.
19. Mizera M, Ostrowicz M, Cielecka-Piontek J. Studies of relationship between Blood-Brain Barrier permeability and chemical structures of drugs with application of deep neutral networks. XXIV EFMC International Symposium on Medicinal Chemistry. August 28 – September 1, 2016, Manchester, United Kingdom, Abstracts Book.
20. Mizera M, Ostrowicz M, Cielecka-Piontek J. Defining a relationship between 3D molecular structures and mutagenicity based on Convolutional Neural Networks as a solution for predicting the toxicity of active pharmaceutical ingredients. Toxicol Lett. 2016;259:S184.
21. Williams AJ, Ekins S, Tkachenko V. Open PHACTS: semantic interoperability for drug discovery. Drug Discov Today. 2012;17(21): 1188-98.
22. Randolph M. Computer-assisted diagnostic decision support: history, challenges, and possible paths forward. Advances in health sciences education 2009;14(1):89-106.
23. Shusaku T. Automated extraction of medical expert system rules from clinical databases based on rough set theory. Inform Sci. 1998;112(1):67-84.
24. Carolyn M. Big data in pharmacy practice: current use, challenges, and the future 2015.
25. Groves P, Kayyali B, Knott D i wsp. The 'big data' revolution in healthcare. McKinsey Quarterly. 2013;2.
26. Evans RS, Abouzelof RH, Taylor C i wsp. Computer surveillance of hospital-acquired infections: a 25 year update. AMIA Annual Symposium Proceedings. Vol. 2009. American Medical Informatics Association 2009.