Pseudonimizacja i anonimizacja danych osobowych w badaniach naukowych – wybrane zagadnienia

*Magdalena Kogut-Czarkowska**

Truizmem jest stwierdzenie, że współczesne badania naukowe, zwłaszcza w obszarach medycznym, społecznym, ekonomicznym wymagają dużych ilości danych osobowych[1]. W szczególności projekty w obszarze ochrony zdrowia korzystają z danych pacjentów. Możliwość zbierania nowych informacji, ale także dostępność do uprzednio zebranych danych jest kluczowa dla prowadzenia badań epidemiologicznych, medycznych i zdrowotnych w oparciu o analitykę dużych zbiorów (*big data*)[2].
Przy powyższych projektach pojawiają się pytania o zgodność badań z obowiązującymi przepisami o ochronie danych osobowych (RODO) i wpływu ograniczeń prawnych na zakres badań. Poszanowanie zasad prywatności jest jedną z najważniejszych zasad wynikających z przepisów Unii Europejskiej[3]. Stąd, pojawiają się napięcia pomiędzy dążeniem do wykorzystania danych do celów naukowych a potrzebą ochrony prywatności osób, których te dane dotyczą. Przykładowo, wskazuje się, że jedną z przeszkód w europejskich badaniach nad epidemią COVID-19 była właśnie kwestia ochrony prywatności pacjentów, a w szczególności brak jasności co do obowiązujących zasad oraz ich zróżnicowanie w zależności od kraju członkowskiego Unii Europejskiej[4].
Celem niniejszego artykułu jest omówienie przepisów odnoszących się do korzystania z danych osobowych do celów badawczych, w szczególności w relacji do badań nad sztuczną inteligencją (AI) w ochronie zdrowia, odwołując się również do doświadczeń związanych z projektem Incisive. Jest to projekt badawczy, wykorzystujący fundusze Unii Europejskiej z programu Horizon 2020 w celu opracowania rozwiązań poprawiających diagnozowanie i przewidywanie rozwoju chorób nowotworowych przy pomocy AI i Big Data[5]. W szczególności, autorka skupi się na kwestii korzystania z danych spseudonimizowanych oraz anonimowych do celów badawczych i niektórych – wynikających z wyboru metody de-identyfikacji danych – konsekwencjach prawnych i praktycznych.

Badania naukowe w RODO

Odniesienie do badań naukowych znajduje się w licznych motywach oraz przepisach RODO[6]. Ogólnym zamierzeniem tych przepisów jest umożliwienie prowadzenia prac naukowych z wykorzystaniem danych osobowych przy jednoczesnym poszanowaniu praw osób, których dane dotyczą. Mówi się wręcz o „uprzywilejowanej pozycji" badań naukowych w RODO[7].

Badania naukowe są rozumiane w RODO w sposób szeroki. Jako przykład takich badań wskazuje się „rozwój technologiczny i demonstrację, badania podstawowe, badania stosowane oraz badania finansowane ze środków prywatnych"[8]. Powyższe ujęcie nie rozróżnia zatem badań prowadzonych dla celów publicznych oraz badań na cele komercyjne. Jednak zgodnie ze wskazaniami Europejskiego Inspektora Danych Osobowych (EIOD)[9] **warunkiem zastosowania szczególnego reżimu RODO dla badań naukowych jest aby były spełnione trzy kryteria**:

1) przetwarzane były dane osobowe;
2) miały zastosowanie odpowiednie sektorowe standardy metodologii i etyki, w tym pojęcia świadomej zgody, rozliczalności i nadzoru oraz
3) badania były prowadzone w celu poszerzenia zbiorowej wiedzy społeczeństwa, a nie służyły przede wszystkim interesom komercyjnym lub prywatnym.

Cele szczególnego reżimu dla badań naukowych oraz przesłanki jakimi kierował się ustawodawca unijny ustanawiając szcze-

* Autorka jest radcą prawnym w kancelarii Timelex w Brukseli, specjalizującym się w prawie danych osobowych, IT oraz IP. Uczestnik projektów badawczych finansowanych przez Unię Europejską w zakresie ochrony danych, nowych technologii, e-zdrowia i sztucznej inteligencji.

[1] Raport Centre for Information Policy Leadership: *Delivering Sustainable AI Accountability in Practice. First Report: Artificial Intelligence and Data Protection in Tension* (2018), s. 12–13; https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_ai_first_report_-_artificial_intelligence_and_data_protection_in_te....pdf (dostęp: 25.5.2021 r.).

[2] *R. Pierce*, Machine Learning for Diagnosis and Treatment: Gymnastics for the GDPR, „European Data Protection Law Review" Nr 3/2018, s. 333.

[3] Art. 8 ust. 1 Karty praw podstawowych Unii Europejskiej, Dz.Urz. C Nr 326 z 26.10.2012 r., s. 391–407 oraz art. 16 ust. 1 Traktatu o funkcjonowaniu Unii Europejskiej (TFUE), Dz.Urz. C Nr 326 z 26.10.2012 r., s. 47–390, stanowią, że każda osoba ma prawo do ochrony danych osobowych jej dotyczących.

[4] *R. Becker*, *A. Thorogood*, *J. Ordish*, *M.J.S. Beauvais*, COVID-19 Research: Navigating the European General Data Protection Regulation. J Med Internet Res. 2020;22(8):e19799. Published 2020 Aug 27. doi:10.2196/19799, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7470233/ (dostęp: 25.5.2021 r.).

[5] Więcej o projekcie na stronie: https://incisive-project.eu/.

[6] Odniesienie znajduje się w art. 5.1b, art. 5.1e, art. 9.2j, art. 14.5b, art. 17.3d, art. 21.6, art. 89 oraz motywach 26, 33, 50, 52, 53, 62, 65, 113, 156, 157, 159, 160, 161 i 162 RODO.

[7] Wstępna opinia EIOD dotycząca ochrony danych i badań naukowych (EDPS A Preliminary Opinion on data protection and scientific research), https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf, s. 18.

[8] Motyw 159 RODO.

[9] Wstępna opinia EIOD dotycząca ochrony danych i badań naukowych, *op. cit.*, s. 12.

gólne zasady wskazane są w motywach preambuły RODO. Motyw 157 wskazuje, że dla ułatwienia badań naukowych dopuszcza się przetwarzanie danych osobowych do celów badań naukowych, z zastrzeżeniem odpowiednich warunków i zabezpieczeń przewidzianych w prawie Unii lub w prawie państwa członkowskiego. Motyw ten wskazuje również na korzyści związane z możliwością łączenia danych z wielu rejestrów. Wskazówki te powinny wskazać kierunek interpretacji przepisów RODO w odniesieniu do zasad przetwarzania danych osobowych w celach prowadzenia badań naukowych.

Dane osobowe. Identyfikatory pośrednie i bezpośrednie

W projektach badawczych punktem wyjścia powinno być ustalenie jakie informacje pochodzące od osób fizycznych lub ich dotyczące są potrzebne do opracowania danego zagadnienia naukowego, a następnie w jakiej formie będą wykorzystywane. Zespół badawczy jest zainteresowany na ogół obserwacją pewnych prawidłowości występujących w ramach grupy osób badanych. Zatem, dla osiągnięcia celu naukowego nie jest potrzebne, aby określony zestaw danych odnosił się do konkretnej, zidentyfikowanej osoby. Przykładowo, jednym z celów projektu INCISIVE, o którym mowa powyżej, jest zbudowanie bazy badań obrazowych (KT, RM, RTG, mammografii i innych) pacjentów chorych na określone rodzaje nowotworów w celu opracowania algorytmu AI, który nauczy się rozpoznawać zmiany chorobowe na zdjęciach, pomagając lekarzowi postawić diagnozę.

Dla treningu algorytmu nie jest potrzebna identyfikacja pacjenta, od którego pochodzi zdjęcie. Również badacz opracowujący algorytm nie musi wiedzieć, że wyniki badania dotyczą osoby o danym imieniu i nazwisku. Dla osiągnięcia celu naukowego wystarczające jest posługiwanie się **danymi „zdezidentyfikowanymi" (*de-identified*)**. To pojęcie jest używane w środowisku naukowym oraz w literaturze[10] w odniesieniu do różnych technik pseudonimizacji oraz anonimizacji łącznie. Warto jednak pamiętać, że ten termin nie występuje w RODO. Samo zatem stwierdzenie, że projekt będzie posługiwał się informacjami „zdezidentyfikowanymi" nie jest wystarczające dla ustalenia zasad ochrony prywatności i zapewnienia zgodności projektu naukowego z tymi zasadami. Konieczne jest ustalenie, czy projekt naukowy będzie wykorzystywał dane osobowe w rozumieniu RODO.

Zgodnie z definicją zawartą w art. 4 pkt 1 RODO **„dane osobowe"** oznaczają wszelkie informacje o zidentyfikowanej lub możliwej do zidentyfikowania osobie fizycznej („osobie, której dane dotyczą"). Z kolei **możliwa do zidentyfikowania osoba fizyczna**, to osoba, którą można bezpośrednio lub pośrednio zidentyfikować, w szczególności na podstawie identyfikatora takiego jak: imię i nazwisko, numer identyfikacyjny, dane o lokalizacji, identyfikator internetowy lub jeden bądź kilka szczególnych czynników określających fizyczną, fizjologiczną, genetyczną, psychiczną, ekonomiczną, kulturową lub społeczną tożsamość osoby fizycznej.

W praktyce często spotyka się wątpliwości, jakie informacje stanowią dane osobowe. O ile takie cechy osoby jak jej imię i nazwisko, PESEL czy numer dokumentu tożsamości łatwo jest zakwalifikować jednoznacznie do danych osobowych, to ocena bardziej ogólnych informacji związanych z daną osobą – np. daty urodzin – jest mniej oczywista. W literaturze[11] spotyka się pojęcia „identyfikatorów bezpośrednich" oraz „identyfikatorów pośrednich", które to rozróżnienie jest przydatne do analizy danej cechy lub ich zbioru w celu określenia czy stanowią dane osobowe. **Identyfikatory bezpośrednie** to konkretne informacje, które odwołują się do osoby fizycznej, takie jak jej imię i nazwisko lub numer identyfikacyjny. Natomiast **identyfikatory pośrednie** (nazywany również *quasi*-identyfikatorami[12]) to dowolne informacje (np. położenie geograficzne w określonym momencie lub opinia na określony temat), które mogą zostać wykorzystane indywidualnie lub w połączeniu z innymi *quasi*-identyfikatorami, przez kogoś, kto ma wiedzę o danej osobie w celu jej ponownej identyfikacji w zbiorze danych. Wśród przykładów identyfikatorów pośrednich podaje się wiek, płeć, wykształcenie, zawód, status społeczno-ekonomiczny, skład gospodarstwa domowego, dochód, stan cywilny, język ojczysty, pochodzenie etniczne, miejsce pracy lub nauki, kod pocztowy czy miejsce zamieszkania. Lista nie jest oczywiście zamknięta. Dowiedziono, że kombinacja kilku z takich pozornie ogólnych cech potrafi prowadzić do identyfikacji osoby, której one dotyczą[13].

Dane anonimowe – czy anonimowość jest możliwa do osiągnięcia?

RODO nie definiuje **pojęcia anonimizacji**, ani **danych anonimowych**. Jednak **motyw 26 rozporządzenia** stanowi, że: „zasady ochrony danych nie powinny (…) mieć zastosowania do informacji anonimowych, czyli informacji, które nie wiążą się ze zidentyfikowaną lub możliwą do zidentyfikowania osobą fizyczną, ani do danych osobowych zanonimizowanych w taki sposób, że osób, których dane dotyczą, w ogóle nie można zidentyfikować lub już nie można zidentyfikować".

Zgodnie natomiast z **definicją ISO**[14], „anonimizacja danych" to proces, w którym dane osobowe są nieodwracalnie zmieniane w taki sposób, że podmiot danych nie może być już zidentyfikowany bezpośrednio lub pośrednio, zarówno przez samego administratora danych, jak i we współpracy z inną stroną.

Anonimizacja może nastąpić poprzez zastosowanie różnych technik i narzędzi stosowanych w celu osiągnięcia anonimowo-

[10] *M. Hintze, K. El Emam*, Comparing the Benefits of Pseudonymization and Anonymization Under the GDPR, „Journal of Data Protection & Privacy" Nr 2(2)/2018, s. 145–158, s. 3.

[11] *M. Hintze, K. El Emam*, Comparing the Benefits…, *op. cit.*, s. 3, a także np. „Wspólny dokument AEPD i EIOD w sprawie 10 nieporozumień związanych z anonimizacją" (10 Misunderstandings Related To Anonymisation), https://edps.europa.eu/data-protection/our-work/publications/papers/aepd-edps-joint-paper-10-misunderstandings-related_en (dostęp: 25.5.2021 r.).

[12] *K. El Emam, B. Malin*, Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk, „Washington D.C.: National Academies Press", 2015, http://www.ncbi.nlm.nih.gov/books/NBK285994.

[13] *D. Barth-Jones*, The 're-identification' of Governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. Then and Now (July 2012), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2076397 oraz *S. Murthy, A. Abu Bakar, F. Abdul Rahim, R. Ramli*, A Comparative Study of Data Anonymization Techniques, 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), 2019, s. 306–309, doi: 10.1109/BigDataSecurity-HPSC-IDS.2019.00063.

[14] ISO/TC 215, Health informatics, https://www.iso.org/obp/ui/#iso:std:iso:25237:ed-1:v1:en.

ści. Dawna Grupa Robocza Art. 29 wskazywała w swojej opinii[15], na dwa ogólne podejścia do anonimizacji: **randomizację** oraz **uogólnianie**. Pierwsze z nich to techniki takie jak: dodawanie zakłóceń, permutacja, prywatność różnicowa. Natomiast do uogólniania należą: agregacja i k-anonimizacja, L-dywersyfikacja/t-bliskość.

Strategia anonimizacji i dobór techniki jest zależny od wielu czynników, m.in. takich jak ilość danych w zbiorze, kategorie osób, typy danych, potencjalne konsekwencje ujawnienia danych dla osób, których dane dotyczą, dostępne zasoby oraz wymagania projektu badawczego. Jednak w zależności od kontekstu lub charakteru danych, ryzyko ponownej identyfikacji nie może być w wystarczającym stopniu ograniczone przez daną technikę. Przykładem jest zbyt mały zbiór danych lub gdy zbiór danych zawiera zbyt dużą liczbę atrybutów demograficznych bądź też dane lokalizacyjne.

W niektórych publikacjach, w tym wytycznych Grupy Roboczej Art. 29 z 2014 r. spotyka się stwierdzenie, że aby dane można było uznać za anonimowe, **anonimizacja** musi być **nieodwracalna**[16]. W nauce podnoszone jest jednak, że **całkowicie anonimowe dane nie istnieją**. Wręcz, część badaczy wskazuje na to, że podejście „zero-jedynkowe" w odniesieniu do danych osobowych i nie-osobowych (anonimowych) nie jest właściwe. Argumentują oni, że anonimizację należy raczej postrzegać jako działanie w zakresie zarządzania ryzykiem[17], które ocenia możliwość zidentyfikowania, należycie uwzględniając wszystkie środki, które mogą być użyte, takie jak czas, zasoby i dostępną technologię. Również w najnowszych publikacjach organów zajmujących się ochroną danych wskazuje się, że wyrażenie „dane anonimowe" nie oznacza, że zbiory danych można po prostu oznaczyć jako anonimowe lub nie[18]. Dlatego też wskazuje się, że nacisk należy położyć na opracowanie procedury, na podstawie której można osiągnąć taki rezultat wobec zbioru danych, w którym **bez znacznego wysiłku** nie da się zidentyfikować poszczególnych osób, również z uwzględnieniem innych niż będących z zbiorze badawczym informacji, które mogą być dostępne stronom chcącym dokonać re-identyfikacji osób fizycznych[19].

Jednym z najbardziej znanych przykładów błędnego przekonania o tym, że udostępnione są dane anonimowe, był eksperyment przeprowadzony na danych składających się z prawie 500 000 ocen filmów nadanych przez użytkowników. W założeniu anonimowe oceny zostały udostępnione publicznie przez serwis streamingowy Netflix[20]. W oparciu o opublikowane informacje, naukowcy przeprowadzili eksperyment wykazujący, że osoba, która posiada nawet niezbyt dużą wiedzę na temat określonego użytkownika serwisu Netflix, może z powodzeniem zidentyfikować jakie oceny nadał poszczególnymi filmom, porównując dane serwisu z innymi dostępnymi publicznie bazami. Co więcej, z dostępnych w ten sposób informacji można było wydedukować preferencje polityczne i inne potencjalnie wrażliwe informacje o wybranych użytkownikach.

Wreszcie, nie można pominąć, że **proces anonimizacji nie pozostaje bez wpływu na ich użyteczność do celów badawczych**[21]. Anonimizacja w kontekście projektów naukowych powinna dążyć do znalezienia właściwej równowagi między ograniczeniem ryzyka ponownej identyfikacji a zachowaniem użyteczności zbioru danych do przewidzianego celu lub celów. Na przykład grupowanie dat urodzenia w odstępach rocznych zmniejszy ryzyko ponownej identyfikacji, ale jednocześnie w niektórych przypadkach zmniejszy użyteczność zbioru danych. Nie oznacza to, że anonimowe dane staną się bezużyteczne, ale raczej, że ich użyteczność będzie zależeć od celu i dopuszczalnego ryzyka ponownej identyfikacji. W szczególności, anonimizacja danych dotyczących zdrowia może potencjalnie zmniejszyć ich przydatność do badań; na przykład anonimizacja danych w badaniach epidemiologicznych może wymagać kilku etapów, w których różne zmienne dotyczące pacjenta są kategoryzowane lub eliminowane (zamazywane), co może prowadzić do obniżenia wartości takich danych.

Pseudonimizacja danych zgodnie z RODO

W przeciwieństwie do danych anonimowych, które nie podlegają RODO (o czym będzie mowa poniżej), dane spseudonimizowane są podkategorią danych osobowych. Zgodnie bowiem z art. 4 pkt 5 RODO „**pseudonimizacja**" oznacza przetworzenie danych osobowych w taki sposób, by nie można ich było już przypisać konkretnej osobie, której dane dotyczą, bez użycia dodatkowych informacji, pod warunkiem że takie dodatkowe informacje są przechowywane osobno i są objęte środkami technicznymi i organizacyjnymi uniemożliwiającymi ich przypisanie zidentyfikowanej lub możliwej do zidentyfikowania osobie fizycznej.

Pseudonimizacja odnosi się zatem do usunięcia lub zastąpienia identyfikatorów (takich jak np. imię i nazwisko, numer PESEL itp.) pseudonimami lub kodami, które są przechowywane oddzielnie i chronione środkami technicznymi i organizacyjnymi. Dane pozostają spseudonimizowane tak długo, jak długo istnieją dodatkowe informacje identyfikujące. Prawidłowo dokonana pseudonimizacja oznacza, że nie można ponownie zidentyfikować osoby na podstawie samych danych opatrzonych

[15] Grupa Robocza Art. 29 Opinia 05/2014 w sprawie technik anonimizacji, s. 3 i n., https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_pl.pdf.

[16] Grupa Robocza Art. 29 Opinia 05/2014 w sprawie technik anonimizacji, *op. cit.*, s. 6 i n.

[17] *M. Finck, F. Pallas*, They Who Must Not Be Identified – Distinguishing Personal from Non-Personal Data Under the GDPR (October 1, 2019). Forthcoming, „International Data Privacy Law", 2020, „Max Planck Institute for Innovation & Competition Research Paper" Nr 19-14, s. 6 i n., https://ssrn.com/abstract=3462948 (dostęp: 25.5.2021 r.).

[18] Wspólny dokument AEPD i EIOD w sprawie 10 nieporozumień związanych z anonimizacją (10 Misunderstandings Related To Anonymisation), https://edps.europa.eu/data-protection/our-work/publications/papers/aepd-edps-joint-paper-10-misunderstandings-related_en.

[19] Przykładowy opis procedury anonimizacji danych [w:] *P.J. Thoral, J.M. Peppink, R.H. Driessen, E.J.G. Sijbrands, E.J.O. Kompanje, L. Kaplan, H. Bailey, J. Kesecioglu, M. Cecconi, M. Churpek, G. Clermont, M. van der Schaar, A. Ercole, A.R.J. Girbes, P.W.G. Elbers,* Amsterdam University Medical Centers Database (AmsterdamUMCdb) Collaborators and the SCCM/ESICM Joint Data Science Task Force. Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example. Crit Care Med. 2021 Jun 1;49(6):e563-e577. doi: 10.1097/CCM.0000000000004916. PMID: 33625129; PMCID: PMC8132908.

[20] *A. Narayanan, V. Shmatikov*, Robust De-anonymization of Large Sparse Datasets, 2008 IEEE Symposium on Security and Privacy (sp 2008), 2008, s. 111–125, doi: 10.1109/SP.2008.33, https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf (dostęp: 25.5.2021 r.).

[21] *P.A. Bonatti, S. Kirrane*, Big Data and Analytics in the Age of the GDPR, 2019 IEEE International Congress on Big Data (BigDataCongress), 2019, s. 7–16, doi: 10.1109/BigDataCongress.2019.00015.

pseudonimem bez dodatkowych, odrębnych informacji. Jednak w odróżnieniu od anonimizacji, jeśli nastąpi pseudonimizacja, ponowna identyfikacja osoby, której dane dotyczą jest możliwa, choć wymaga spełnienia dodatkowych warunków.

W kontekście projektów badawczych, oznacza to, że zespół badawczy (lub inna zaufana strona) powinna dysponować „**kluczem**", który – w określonych sytuacjach – może być użyty do połączenia danych używanych w badaniu z konkretnymi uczestnikami. Taka osoba może zatem ponownie zidentyfikować osoby, których dane dotyczą. Informacje na temat pierwotnych wartości i technik stosowanych do tworzenia pseudonimów muszą być jednak przechowywane w sposób organizacyjnie i technicznie oddzielony od danych opatrzonych pseudonimem[22].

Warto podkreślić, że dane nie są prawidłowo spseudonimizowane, jeżeli konkretną osobę, której dane dotyczą, można zidentyfikować na podstawie samych danych, bez dodatkowych informacji („klucza"). Taka sytuacja może mieć miejsce, gdy identyfikatory pośrednie w połączeniu z dostępnymi także publicznie informacjami umożliwiają identyfikację osoby, chociażby administrator przechowywał identyfikatory bezpośrednie (np. nazwiska) oddzielnie i w bezpieczny sposób. Pseudonimizacja jest również nieskuteczna, jeśli osoba z zewnątrz jest w stanie określić oryginalne wartości na podstawie pseudonimów.

Istnieją różne techniki pseudonimizacji danych. Grupa Robocza Art. 29 wymieniła najczęściej stosowane **techniki pseudonimizacji**[23]:

1) **szyfrowanie z kluczem tajnym**: w tym przypadku posiadacz klucza może z łatwością ponownie zidentyfikować każdą osobę, której dane dotyczą, poprzez odszyfrowanie zbioru danych, ponieważ dane osobowe nadal znajdują się w tym zbiorze danych, chociaż w zaszyfrowanej formie;
2) **funkcja skrótu (*hash function*)**: oznacza funkcję, która z wkładu każdej wielkości daje wynik stałej wielkości i której nie można odwrócić; oznacza to, że ryzyko odwrócenia, występujące przy szyfrowaniu, już nie istnieje. Funkcje skrótu są zwykle opracowane w taki sposób, aby można było prowadzić stosunkowo szybkie obliczenia, stąd mogą być podatne na ataki siłowe. Stosowanie funkcji skrótu z losowym ciągiem znaków (w której do skracanego atrybutu dodaje się losowy ciąg znaków, ang. *salt*) może ograniczyć prawdopodobieństwo skutecznego ataku;
3) **funkcja skrótu z kluczem, w przypadku której klucz jest przechowywany**: oznacza określoną funkcję skrótu, która wykorzystuje klucz tajny jako dodatkowy wkład. To powoduje, że atakującemu trudniej jest odtworzyć funkcję bez znajomości tajnego klucza;
4) **szyfrowanie deterministyczne lub funkcja skrótu z kluczem, w przypadku której klucz jest usuwany**: technika ta może być utożsamiona z wybieraniem losowego numeru jako pseudonimu dla każdego atrybutu w bazie danych, a następnie z usuwaniem tabeli korelacji;
5) **tokenizacja**: technika zwykle polega na stosowaniu mechanizmów szyfrowania jednokierunkowego lub na przypisaniu, za pomocą funkcji indeksu, sekwencji liczb lub losowo wygenerowanych liczb, które nie zostały w sposób matematyczny uzyskane z danych pierwotnych. Opiera się ona na omówionych wcześniej technikach i ta jest często stosowana w sektorze finansowym.

Kodowanie i podwójne kodowanie są często stosowanymi środkami bezpieczeństwa danych w projektach badań na danych medycznych. Tym niemniej nawet szyfrowanie danych nie prowadzi do ich anonimizacji, a dane nie staną się anonimowe, nawet jeśli klucz deszyfrujący zostanie zakodowany dwukrotnie (podwójne kodowanie).

Konsekwencje wyboru sposobu de-identyfikacji danych

Mimo że RODO zawiera definicję danych spseudonimizowanych, jak też – poprzez motyw 26 – precyzuje jakie informacje na gruncie przepisów powinny być uznawane za dane anonimowe, w środowisku naukowym nadal istnieje wiele wątpliwości w tym zakresie. W literaturze[24] wskazuje się, że jednym ze źródeł tego niezrozumienia mogą być **względy historyczne**. W szczególności zgodnie z przyjętą w Stanach Zjednoczonych w 1996 r. ustawą o przenoszalności i odpowiedzialności w ubezpieczeniach zdrowotnych (*US Health Insurance Portability and Accountability Act of 1996*, HIPAA) możliwe było uznanie za „dane anonimowe" lub „zdezidentyfikowane" (*deidentified*) danych zakodowanych za pomocą klucza, o ile zostały spełnione określone warunki[25]. Również wytyczne dotyczące anonimizacji[26] wydane przez biuro komisarza ds. informacji (ICO) w Wielkiej Brytanii w 2012 r. (na gruncie nieobowiązującej już dyrektywy 95/46/WE[27]), niekiedy wymiennie posługiwały się pojęciami danych anonimowych oraz danych zakodowanych[28]. Wśród innych przykładów zapisów traktujących dane zakodowane jako dane nie-osobowe można wymienić postanowienia programu EU-US Privacy Shield[29].

[22] Więcej o prawidłowych technikach pseudonimizacji ENISA w rekomendacjach: Recommendations on shaping technology according to GDPR provisions: An overview on data pseudonymisation, listopad 2018 r. oraz Data Pseudonymisation: Advanced Techniques and Use Cases, styczeń 2021 r.

[23] Grupa Robocza Art. 29 Opinia 05/2014 w sprawie technik anonimizacji, *op. cit.*, s. 22–23.

[24] *D. Peloquin, M. DiMaio, B. Bierer, et al.,* Disruptive and avoidable: GDPR challenges to secondary research uses of data [w:] European Journal of Human Genetics, Nr 3/2020, https://doi.org/10.1038/s41431-020-0596-x.

[25] Zob. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard.

[26] United Kingdom Information Commissioner's Office, Anonymisation: Managing Data Protection Risk Code of Practice (2012), Annex 2, https://ico.org.uk/media/1061/anonymisation-code.pdf.

[27] Dyrektywa 95/46/WE Parlamentu Europejskiego i Rady z 24.10.1995 r. w sprawie ochrony osób fizycznych w zakresie przetwarzania danych osobowych i swobodnego przepływu tych danych, Dz.Urz. L Nr 281 z 23.11.1995 r., s. 31–50.

[28] Wskazuje na to m.in. analiza przykładu anonimizacji danych: „W badaniu klinicznym tylko dane zakodowane kluczem są przekazywane przez kliniki badawcze (pracowników służby zdrowia) do firm farmaceutycznych sponsorujących badania. Nie ujawnia się żadnych danych osobowych. Klucze deszyfrujące są przechowywane w ośrodkach badawczych przez badaczy klinicznych, którzy zgodnie z zasadami dobrej praktyki klinicznej i tajemnicy zawodowej nie mogą ujawniać tożsamości uczestników badania. Sponsorzy badań mogą udostępniać zakodowane kluczami dane oddziałom za granicą, współpracownikom naukowym, i organom regulacyjnym na całym świecie. We wszystkich przypadkach, jednak odbiorcy danych są zobowiązani do zachowania poufności oraz przestrzegania ograniczeń dotyczących ponownego wykorzystania i ponownej identyfikacji, zarówno nałożonych umową, jak też wymaganych prawem. Biorąc pod uwagę te zabezpieczenia, ryzyko ponownej identyfikacji danych kodowanych za pomocą klucza ujawnionych przez sponsora stronie trzeciej w ramach takich zobowiązań jest niezwykle niskie".

[29] *D. Peloquin, M. DiMaio, B. Bierer, et al.,* Disruptive and avoidable…, *op. cit.*

Powyższe czynniki, jak również fakt przenikania się środowisk naukowych z wielu krajów i konieczność prowadzenia międzynarodowych projektów w zgodzie z wieloma, często będącymi w konflikcie, zestawami przepisów, niewątpliwie mają wpływ na obecne niejasności dotyczące rozumienia pojęć „anonimizacji", „pseudonimizacji" oraz szerzej „de-identyfikacji" danych osobowych. Dowodem na trudności jakie powoduje ta kwestia dla środowiska naukowego jest choćby otwarty list wystosowany przez praktyków[30] do władz w Wielkiej Brytanii, mówiący o mieszaniu tych pojęć przez komisje etyczne oraz braku jednoznacznych wytycznych dla badaczy.

Tymczasem, między korzystaniem z danych anonimowych a pseudonimowanymi istnieje zasadnicza różnica w kwestii obowiązków prawnych, jaki ten wybór za sobą pociąga. Poniżej omówię najważniejsze konsekwencje podjęcia decyzji o anonimizacji lub pseudonimizacji danych w kontekście obowiązków wynikających z RODO, mających znaczenie dla projektu badawczego[31].

Anonimizacja danych jako proces przetwarzania danych osobowych

W najprostszym ujęciu, RODO reguluje jedynie przetwarzanie danych osobowych. Zatem rozporządzenie nie dotyczy przetwarzania anonimowych informacji, w tym przetwarzania takich informacji do celów statystycznych lub naukowych. **Dane anonimowe nie podlegają** wobec powyższego **zasadom ochrony danych, podczas gdy dane spseudonimizowane są poddane reżimowi RODO**.

Rodzi to poważne konsekwencje w całym cyklu projektu badawczego. Jeśli w danym projekcie zespół badawczy wykorzystuje **dane anonimowe**, nie musi wykazywać podstawy prawnej przetwarzania danych. Do tego przetwarzania nie mają zastosowania przepisy RODO. Tym niemniej, administrator niezanonimizowanych danych osobowe, stanowiących dane wyjściowe, nadal podlega zasadom ochrony danych osobowych. Musi zatem zbadać, czy istnieje podstawa prawna do procesu przetwarzania danych jakim jest ich anonimizacja[32]. Przykładowo zatem, jeśli dane potrzebne do celów badawczych będą zanonimizowanymi danymi pacjentów pozyskanymi od placówki zdrowotnej, to placówka ta powinna dysponować podstawą prawną do zbierania tych danych, przechowywania ich oraz anonimizacji na potrzeby projektów badawczych. Jednak kwestia podstawy prawnej anonimizacji danych nie powinna w praktyce stanowić przeszkody. Grupa Robocza Art. 29 wyjaśniła bowiem w przeszłości, że przetwarzanie danych osobowych w celu ich całkowitej anonimizacji jest zasadniczo „zgodne z celem, dla którego dane osobowe zostały pierwotnie zebrane"[33] oraz, że w związku z tym nie wymaga dodatkowej podstawy prawnej. Wydaje się, że to stanowisko zachowuje aktualność również na gruncie RODO. Jednocześnie, w celu pełnej transparentności dotyczącej przetwarzania danych wobec osób, których dotyczą, rekomendowane jest, aby podczas gromadzenia danych osobowych informować te osoby (w omówionym przykładzie – pacjentów), iż jednym z celów ich przetwarzania jest anonimizacja ich danych w celu przyszłego wykorzystania w projektach badawczych.

Podstawa prawna przetwarzania danych spseudonimizowanych w projektach badawczych

Tak jak wszystkie inne dane osobowe, również **dane spseudonimizowane mogą być wykorzystywane w badaniach naukowych tylko wtedy, gdy jest to dozwolone przez RODO**. Innymi słowy, organizacja prowadząca projekt badawczy powinna ustalić czy istnieje podstawa prawna umożliwiająca wykorzystanie danych osobowych do celów prowadzenia projektu naukowego.

Punktem wyjścia w takiej analizie powinno być **ustalenie pochodzenia danych i celu ich pierwotnego zebrania**. Podstawowym pytaniem na które zespół badawczy musi odpowiedzieć jest czy podstawowym celem wykorzystania zbieranych danych jest prowadzenie projektu badawczego (pierwotny cel przetwarzania). Alternatywnie projekt badawczy może bowiem opierać się na danych już uprzednio zebranych do innego celu (wtórny cel przetwarzania danych). Rozróżnienie to determinuje wybór właściwej podstawy prawnej do wykorzystywania danych do celów badawczych.

Jeśli projekt badawczy będzie zbierał nowe dane w celu ich analizy w toku badań naukowych, częstą praktyką jest **uzyskiwanie zgody badanego na przetwarzanie jego danych**. Zgoda ta musi być wówczas zebrana zgodnie z art. 6 ust. 1 lit. a) RODO. W przypadku danych wrażliwych, konieczne jest zebranie zgody wyraźnej (art. 9 ust. 2 lit. a) RODO). Zgoda musi spełniać wymogi z art. 4 pkt 11 RODO, co oznacza, że musi być to dobrowolne, konkretne, świadome i jednoznaczne okazanie woli. Jednak, w przypadku badań naukowych, często nie jest możliwe uprzednie ustalenie jaki będzie ostateczny cel badań. Okoliczność ta *prima facie* stoi w sprzeczności z zasadą konkretności zgody. Zostało to wskazane w motywie 33 RODO, w którym ustawodawca sam dostrzega, że w momencie zbierania danych często nie da się w pełni zidentyfikować celu przetwarzania danych osobowych na potrzeby badań naukowych. Aby wyjść naprzeciw tym trudnościom, zbierając zgodę rekomendowane jest, aby osoby, których dane dotyczą, powinny móc ją wyrazić na niektóre obszary badań naukowych. Dodatkowo, wskazany motyw RODO podkreśla, że badania powinny być zgodne z uznanymi normami etycznymi w zakresie badań naukowych.

Co do zasady, **nie jest możliwe zbieranie *in blanco* zgody na korzystanie z danych na cele prowadzonych w przyszłości,**

[30] The real issue with health data regulation: An open letter to UK policymakers, https://www.lexology.com/library/detail.aspx?g=b9a5c73f-36ad-49fe-b58c-4acdbc18b0c3.

[31] Z uwagi na charakter publikacji, której celem jest przede wszystkim uwypuklenie różnic między korzystaniem z danych anonimowych a danych spseudonimizowanych, w kontekście naukowym nie będą szczegółowo omówione wszystkie aspekty wynikające z przetwarzania danych osobowych. W szczególności pominięta zostanie konieczność minimalizacji danych, kwestie zgłaszania naruszeń ochrony danych, DPIA, rejestry przetwarzania, okresy retencji.

[32] Por. dokument Europejskiej Rady Ochrony Danych w sprawie odpowiedzi na wniosek Komisji Europejskiej o wyjaśnienia dotyczące spójnego stosowania GDPR, ze szczególnym uwzględnieniem badań w dziedzinie zdrowia (dostępny w wersji angielskiej: *Document on response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research*), pkt 43, https://edpb.europa.eu/sites/default/files/files/file1/edpb_replyec_questionnaireresearch_final.pdf.

[33] Grupa Robocza Art. 29: opinia 05/2014 w sprawie technik anonimizacji, *op. cit.*, s. 7. Wskazano w niej, że: „Grupa robocza uważa, że anonimizacja jako przykład dalszego przetwarzania danych osobowych może być uznana za zgodną z pierwotnymi celami przetwarzania, ale tylko pod warunkiem, że proces anonimizacji umożliwia wiarygodne uzyskanie zanonimizowanych informacji w rozumieniu opisanym w niniejszym dokumencie".

niezdefiniowanych projektów badawczych. Przedmiot badań musi być określony co najmniej w zarysie. Jednak spotyka się również głosy, że w dobie badań nad sztuczną inteligencją, określenie celu przetwarzania w momencie zbierania zgody jest niemożliwe[34].

Co ciekawe, w będącym obecnie w toku prac parlamentarnych projekcie rozporządzenia unijnego Akt w sprawie zarządzania danymi (*Data Governance Act*)[35], ustawodawca przewiduje, że osoby fizyczne będą mogły w ramach tzw. altruistycznego podejścia do danych, wyrazić zgodę na wykorzystywanie ich danych *nieosobowych* bez żądania wynagrodzenia, do celów realizowanych w interesie ogólnym, takich jak cele badań naukowych lub poprawa jakości usług publicznych[36]. W praktyce rozgraniczenie danych osobowych od danych nieosobowych nastręcza niejednokrotnie dużych trudności. Zatem pozostaje niejasne jak pogodzić proponowany mechanizm z wymogiem uzyskania konkretnej, spełniającej wymogi RODO zgody na korzystanie z danych osobowych, na co zresztą zwrócono uwagę we wspólnym stanowisku EROD i EIOD[37].

Niezależnie od zagadnień prawidłowego formułowania zakresu zgody, ta przesłanka przetwarzania może być problematyczna w projektach naukowych również z innych względów. W szczególności, trudności może powodować **wzajemna relacja zgody na udział w badaniu jako takim oraz zgody na przetwarzanie danych**. Przykładem takiej sytuacji są badania kliniczne. W celu wzięcia udziału w badaniu klinicznym, zgodnie z przepisami prawa farmaceutycznego (a w przyszłości rozporządzenia unijnego o badaniach klinicznych[38]), potrzebna jest tzw. świadoma zgoda[39] pacjenta. Jej wyrażenie nie jest tożsame z udzieleniem zgody na korzystanie z danych osobowych, o której mowa w RODO. Tym niemniej, bez wyrażenia zgody na przetwarzanie danych nie będzie możliwe wzięcie udziału w badaniu klinicznym. Ponadto, w wielu przypadkach uczestnik może nie być w stanie wyrazić zgody na przetwarzanie danych w sposób świadomy i niezależny – na przykład, jeśli od niej będzie zależeć jego stan zdrowia. To rodzi pytanie o dobrowolność zgody na przetwarzanie danych w kontekście badań naukowych[40]. Z tego względu, również w wytycznych różnych organów Unii Europejskiej, rekomendowano inne niż zgoda przesłanki do przetwarzania danych w celach badawczych[41].

Nie można również pominąć, że, jeżeli zgoda stanowi podstawę przetwarzania danych, projekt badawczy musi zapewnić uczestnikom możliwość jej wycofania w dowolnym momencie, zgodnie z zasadą wynikającą z art. 7 ust. 3 RODO. Nie przewidziano bowiem w RODO wyjątków w tym zakresie w odniesieniu do badań naukowych. Zostało to podkreślone również w dokumentach Grupy Roboczej Art. 29, która wskazała, że: „Jeśli administrator otrzyma żądanie wycofania zgody, musi zasadniczo bezzwłocznie usunąć dane osobowe, jeżeli chce w dalszym ciągu wykorzystywać dane do celów badania"[42].

Należy zatem podkreślić, że **zgoda na przetwarzanie danych nie jest jedyną możliwą przesłanką przetwarzania danych w kontekście projektów badawczych**. Potwierdzają również wytyczne Grupy Roboczej Art. 29, która wprost wskazuje, że: „możliwe są inne podstawy prawne, takie jak art. 6 ust. 1 lit. e) lub f), jeżeli istnieją właściwe zabezpieczenia, np. wymogi przewidziane w art. 89 ust. 1, a przetwarzanie jest uczciwe, zgodne z prawem, przejrzyste oraz zgodne z normami w zakresie minimalizacji danych i prawami poszczególnych osób. Dotyczy to również szczególnych kategorii danych zgodnie z odstępstwem, o którym mowa w art. 9 ust. 2 lit. j)"[43]. Jednak w odniesieniu do tego przepisu, z kolei EIOD odwołuje się do kompetencji państw członkowskich, które na mocy art. 9 ust. 4 RODO mogą wprowadzać dalsze warunki, w tym ograniczenia, w odniesieniu do przetwarzania danych genetycznych, danych biometrycznych lub danych dotyczących zdrowia. Zdaniem EIOD jest to nowy obszar, który „wymaga przyjęcia prawa UE lub państwa członkowskiego, zanim wykorzystanie specjalnych kategorii danych do celów badawczych będzie mogło w pełni funkcjonować[44]." Zatem, zakres zastosowania art. 9 ust. 2 lit. j), jako podstawy prawnej do prowadzenia obecnie badań naukowych z wykorzystaniem danych osobowych jest niejasny, w szczególności wobec braku przyjęcia przez wszystkie państwa członkowskie lub Unii Europejskiej jako całości szczegółowych przepisów w tym zakresie.

Jeszcze inaczej należy zapatrywać się na **sytuację, gdy dane, z których badacz chce skorzystać zostały pierwotnie zebrane w innym celu niż prowadzenie projektu naukowego**. Jest to częsta sytuacja, coraz więcej mówi się też o korzyściach i szansach płynących z analizy w celach naukowych tzw. *health big data*, gromadzonych m.in. w dokumentacji medycznej szpitali, prowadzonej w celach leczniczych[45].

Patrząc na tę problematykę, z punktu widzenia RODO, należy zauważyć, że ograniczenie celu jest podstawową zasadą ochrony danych, zgodnie z którą dane są gromadzone do określonych, jednoznacznych i legalnych celów i nie mogą być dalej przetwarzane w sposób niezgodny z tymi celami (art. 5 ust. 1 RODO). Do-

[34] *R. Pierce*, Machine Learning for Diagnosis and Treatment: Gymnastics for the GDPR, „European Data Protection Law Review" Nr 3/2018, s. 339.

[35] Projekt Rozporządzenia Parlamentu Europejskiego i Rady w sprawie europejskiego zarządzania danymi (Akt w sprawie zarządzania danymi), https://eur-lex.europa.eu/legal-content/PL/TXT/HTML/?uri=CELEX:52020PC0767&from=EN (dostęp: 18.5.2021 r.); dalej jako: projekt Aktu w sprawie zarządzania danymi.

[36] Art. 2 pkt 10 projektu Aktu w sprawie zarządzania danymi.

[37] Wspólna opinia EROD i EIOD w sprawie projektu rozporządzenia unijnego akt w sprawie zarządzania danymi (EDPB-EDPS Joint Opinion 03/2021 on the Proposal for a regulation of the European Parliament and of the Council on European data governance, „Data Governance Act"), https://edpb.europa.eu/system/files/2021-03/edpb-edps_joint_opinion_dga_en.pdf (dostęp: 25.5.2021 r.).

[38] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 536/2014 z 16.4.2014 r. w sprawie badań klinicznych produktów leczniczych stosowanych u ludzi oraz uchylenia dyrektywy 2001/20/WE, Dz.Urz. L Nr 158 z 27.4.2014 r., s. 1–76.

[39] Art. 37f ustawy z 6.9.2001 r. – Prawo farmaceutyczne, t. jedn.: Dz.U. z 2021 r. poz. 974, 981.

[40] Kwestią tą zajęła się bliżej EROD w opinii Nr 3/2019 w sprawie pytań i odpowiedzi dotyczących wzajemnych zależności między rozporządzeniem w sprawie badań klinicznych (RBK) a ogólnym rozporządzeniem o ochronie danych (RODO) (art. 70 ust. 1 lit. b).

[41] Więcej na ten temat *G. Verhenneman, K. Claes, J.J. Derèze, P. Herijgers, C. Mathieu, F.E. Rademakers, R. Reyda, M. Vanautgaerden,* How GDPR Enhances Transparency and Fosters Pseudonymisation in Academic Medical Research, „European Journal of Health Law" Nr 27(1)/2020, s. 35–57, doi: https://doi.org/10.1163/15718093-12251009.

[42] Grupa Robocza Art. 29, WP259: Wytyczne dotyczące zgody na mocy rozporządzenia 2016/679, ostatnio zmienione i przyjęte 10.4.2018 r., s. 33.

[43] Grupa Robocza Art. 29, WP259, *op. cit.*, s. 31.

[44] Wstępna opinia EIOD dotycząca ochrony danych i badań naukowych, *op. cit.*, s. 17.

[45] 18 Examples Of Big Data Analytics In Healthcare That Can Save People, https://www.datapine.com/blog/big-data-examples-in-healthcare/.

datkowo, zgodnie z motywem 50 RODO, przetwarzanie danych osobowych do celów innych niż cele, w których dane te zostały pierwotnie zebrane, powinno być dozwolone wyłącznie w przypadkach, gdy jest zgodne z celami, w których dane osobowe zostały pierwotnie zebrane. W takim przypadku nie jest wymagana odrębna podstawa prawna inna niż podstawa prawna, która umożliwiła zbieranie danych osobowych.

Co istotne z uwagi na tematykę niniejszego artykułu, omawiany motyw stanowi jednak, że dalsze przetwarzanie **do celów badań naukowych** lub historycznych albo do celów statystycznych, **powinny być uznawane za operacje przetwarzania zgodne z prawem i z pierwotnymi celami**. Założenie to jest następnie wprowadzone art. 5 ust. 1 lit. b) RODO. Zgodnie z nim, dalsze przetwarzanie do celów archiwalnych w interesie publicznym, do celów badań naukowych lub historycznych albo do celów statystycznych **nie jest uznawane, w myśl art. 89 ust. 1 RODO, za niezgodne z pierwotnymi celami** („ograniczenie celu").

Powyższe przepisy stanowią **szczególny reżim dotyczący korzystania z uprzednio zebranych danych osobowych do celów naukowych**. Cel naukowy jest, w myśl tego reżimu, niejako uprzywilejowany na gruncie RODO. W myśl przytoczonych zasad bowiem istnieje „**domniemanie zgodności**" (*presumption of compatibility*)[46], iż przetwarzanie danych z celach naukowych nie jest uznawane za niezgodne z pierwotnymi celami przetwarzania, a więc nie zachodzi wbrew zasadzie ograniczenia celu. Pojęcie to jest często określane jako „wtórne wykorzystanie" (*secondary use*), „dalsze przetwarzanie" (*further processing*) lub „dalsze wykorzystanie" (*further use*).

Zasady, o których mowa powyżej, nie oznaczają jednak automatycznego uznania, że w każdej sytuacji będzie możliwe przetwarzanie wcześniej zebranych danych do celów naukowych. W literaturze zaleca się przeprowadzenie „testu kompatybilności" celów pierwotnego oraz badawczego. Ocena powinna zostać dokonana w oparciu o związek między tymi celami, kontekst przetwarzania, charakter danych, których dotyczy przetwarzanie, możliwe konsekwencje przetwarzania oraz istnienie odpowiednich zabezpieczeń (art. 6 ust. 4 RODO)[47]. Ponadto, aby móc korzystać z domniemania zgodności, należy spełnić szereg kryteriów oraz podlegać odpowiednim zabezpieczeniom organizacyjnym i technicznym, co jest w doktrynie[48] uważane za „centralny element" przepisów dotyczący przetwarzania danych dla celów badawczych. Projekt badawczy musi zatem wykazać zgodność z art. 89 ust. 1 RODO.

Zgodnie z art. 89 RODO przetwarzanie do celów archiwalnych w interesie publicznym, do celów badań naukowych lub historycznych albo do celów statystycznych podlega **odpowiednim zabezpieczeniom dla praw i wolności osoby**, której dane dotyczą, zgodnie z niniejszym rozporządzeniem. Zabezpieczenia te polegają na wdrożeniu środków technicznych i organizacyjnych zapewniających poszanowanie zasady minimalizacji danych. Środki te mogą też obejmować **pseudonimizację** danych, o ile pozwala ona realizować powyższe cele. Jeżeli cele te można zrealizować w drodze dalszego przetwarzania danych, które nie pozwalają albo przestały pozwalać na zidentyfikowanie osoby, której dane dotyczą, cele należy realizować w ten sposób. Oprócz pseudonimizacji danych, inne przykłady wdrożonych środków mogą obejmować zapewnienie bezpieczeństwa informatycznego, a także dokonanie oceny skutków dla ochrony danych dla projektu badawczego, zgodnej z wymogiem art. 35 RODO. W kontekście projektów badawczych, takich jak INCISIVE, przykładami środków zabezpieczeń może być uzyskanie wymaganych zgód komisji etycznych, a na etapie prac badawczych dodatkowo ograniczenie wglądu do danych do określonego grona naukowców oraz rejestrowanie każdorazowego dostępu do danych[49].

Podsumowując, na podstawie wspomnianego art. 5 ust. 1 lit. b) RODO dalsze przetwarzanie danych osobowych w ramach projektów badawczych powinno być *prima facie* uznane za zgodne z pierwotnymi celami. Należy też zauważyć, że zgodnie z powołanym powyżej motywem 50 RODO, jeżeli porównywane cele przetwarzania nie są niezgodne, nie ma konieczności szukania nowej podstawy prawnej do nowego celu przetwarzania[50]. Przepisy krajowe mogą jednak nakładać dodatkowe warunki, w szczególności w odniesieniu do danych o zdrowiu (na podstawie wspomnianego art. 9 ust. 4 RODO). Z uwagi na skomplikowanie tego reżimu i wątpliwości wyrażane przez niektórych przedstawicieli środowiska, EROD zapowiedziała przygotowanie dokumentu szerzej analizującego zagadnienie wykorzystywania danych osobowych do celów badawczych[51].

Na koniec warto zauważyć, że z praktycznego punktu widzenia, istnienie szczególnego reżimu wtórnego wykorzystania danych do celów naukowych jest pomocne w sytuacjach, gdy w praktyce zwrócenie się do podmiotów danych o zgodę na wykorzystanie ich danych jest niemożliwe. Może być tak z różnych przyczyn, takich jak brak kontaktu z osobami, których dane dotyczą, trudności i koszty w organizacji ponownego zebrania zgód, czy też wreszcie fakt, że w przypadku, gdy część osób z oryginalnego zbioru odmówi udzielenia zgody, będzie miało to wpływ na jakość oraz zawartość zbioru danych. Ograniczenie mechanizmu wtórnego wykorzystywania danych osobowych do celów badawczych może przełożyć się na zablokowanie możliwości prowadzenia innowacyjnych projektów badawczych w UE w oparciu o uprzednio zebrane dane.

---

[46] Terminem „presumption of compatibility" posługuje się EOID we wstępnej opinii dotyczącej ochrony danych i badań naukowych.

[47] *G. Verhenneman, K. Claes, J.J. Derèze, P. Herijgers, C. Mathieu, F.W. Rademakers, R. Reyda, M. Vanautgaerden,* How GDPR Enhances Transparency…, *op. cit.*, s. 40.

[48] *G. Chassang,* The impact of the EU general data protection regulation on scientific research, Ecancermedicalscience. 2017 Jan 3;11:709. doi: 10.3332/ecancer.2017.709. PMID: 28144283; PMCID: PMC5243137, https://ecancer.org/en/journal/article/709-the-impact-of-the-eu-general-data-protection-regulation-on-scientific-research.

[49] *M. Shabani, G. Chassang, L. Marelli,* „The Impact of the GDPR on the Governance of Biobank Research" [w:] GDPR and Biobanking, Law, Governance and Technology Series book series (LGTS, v. 43), s. 45–60.

[50] Tak również *G. Verhenneman, K. Claes, J.J. Derèze, P. Herijgers, C. Mathieu, F.E. Rademakers, R. Reyda, M. Vanautgaerden,* How GDPR Enhances Transparency…, *op. cit.*, s. 48.

[51] Dokument Europejskiej Rady Ochrony Danych w sprawie odpowiedzi na wniosek Komisji Europejskiej o wyjaśnienia dotyczące spójnego stosowania GDPR, ze szczególnym uwzględnieniem badań w dziedzinie zdrowia (dostępny w wersji angielskiej: Document on response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research), pkt 3, https://edpb.europa.eu/sites/default/files/files/file1/edpb_replyec_questionnaireresearch_final.pdf. EROD prowadziła również publiczne konsultacje https://edpb.europa.eu/news/news/2021/edpb-stakeholder-event-processing-personal-data-scientific-research-purposes_en.

Bezpieczeństwo danych

Obok podstawy prawnej, kluczowym czynnikiem, który należy rozważyć podejmując badania jest **zapewnienie bezpieczeństwa danych**.

W przypadku danych anonimowych RODO nie będzie miało zastosowania w zakresie obowiązków i standardu ochrony tych informacji. Jednak, jak wykazano powyżej, przeprowadzenie prawidłowej anonimizacji danych osobowych jest wyzwaniem od strony technicznej i powinno być poprzedzone gruntowną analizą zarówno zakresu danych oraz ryzyka, które mogłoby się ziścić, gdyby dane zostały „zdeanonimizowane". Dopiero na tej podstawie można przejść do kolejnego etapu, oznaczającego opracowanie metody oraz procesu anonimizacji. Dobranie metody anonimizacji zależy od rozważenia szeregu czynników. Nie ma rozwiązań, które zawsze pasują do każdego zastawu danych. Niektóre organy nadzorcze przygotowały wytyczne, zalecające przygotowanie planu anonimizacji, dokumentującego wszystkie istotne techniki i procesy anonimizacji wraz z ich uzasadnieniem[52].

Z kolei, **w przypadku gdy projekt korzysta z danych spseudonimizowanych, dane te powinny być zabezpieczone tak jak dane osobowe**, zgodnie z art. 32 RODO, poprzez wdrożenie odpowiednich środków technicznych i organizacyjnych w celu zapewnienia poziomu bezpieczeństwa odpowiedniego do ryzyka, w tym m.in. odpowiednio:

1) szyfrowanie danych osobowych;
2) zdolność do zapewnienia ciągłej poufności, integralności, dostępności i odporności systemów i usług przetwarzania;
3) zdolność do przywrócenia dostępności i dostępu do danych osobowych w odpowiednim czasie w przypadku incydentu fizycznego lub technicznego oraz
4) proces regularnego testowania, oceny i ewaluacji skuteczności środków technicznych i organizacyjnych mających na celu zapewnienie bezpieczeństwa przetwarzania danych. Jednocześnie, pseudonimizacja może pomóc spełnić zobowiązania w zakresie ochrony danych, w szczególności zasady minimalizacji danych i ograniczenia przechowywania (art. 5 ust. 1 lit. c) i art. 5 ust. 1 lit. e) RODO, gdyż jej prawidłowe przeprowadzenie powinno pośrednio prowadzić przez eliminację informacji zbędnych do osiągnięcia celów projektu.

Dodatkowo, przy założeniu prawidłowej pseudonimizacji, posługiwanie się w toku projektu badawczego danymi zakodowanymi wiąże się z mniejszym ryzykiem dla osób niż używanie danych w pełni identyfikujących osoby fizyczne. Zatem, pseudonimizacja może być stosowana jako jeden ze środków bezpieczeństwa. Przepis art. 32 ust. 1 lit. a) RODO, wymienia pseudonimizację jako jeden z „odpowiednich technicznych środków organizacji", a art. 6 ust. 4 odnosi się do odpowiednich środków bezpieczeństwa, które mogą obejmować szyfrowanie lub pseudonimizację[53].

Transfery danych osobowych poza Europejski Obszar Gospodarczy

Kolejnym zagadnieniem istotnym dla projektów badawczych jest **ustalenie zasad dostępu do danych i ich przekazywania**. Nierzadko, projekty prowadzone są przez międzynarodowe konsorcja badawcze, bądź też korzystają w zasobów przechowywania danych, które mogą być zlokalizowane poza UE. Stąd ustalenie czy mają zastosowanie zasady transferów danych osobowych ma szczególne znaczenie.

Przekazywanie danych anonimowych za granicę nie wymaga spełnienia dodatkowych wymogów, z punktu widzenia RODO. Odwrotnie jest w przypadku danych spseudonimizowanych. Przekazywanie takich danych poza Europejski Obszar Gospodarczy wymaga zastosowania dodatkowych zabezpieczeń, jeżeli kraj przeznaczenia nie jest objęty decyzją Komisji Europejskiej stwierdzającą odpowiedni poziom ochrony.

Nie pozostaje to bez negatywnych konsekwencji dla projektów badawczych. Oszacowano, że konieczność dostosowania się do zasad RODO wpłynęła w samym 2019 r. na ponad 5000 projektów współpracy (projekty z udziałem Narodowych Instytutów Zdrowia Stanów Zjednoczonych i krajów EOG)[54]. Co więcej, środowiska naukowe wskazują, że w wyniku wyroku Trybunału Sprawiedliwości w sprawie *Schrems II*[55], który rygorystycznie określa zasady postępowania w przypadku transferów danych poza EOG, dostęp do danych z innych krajów niż w UE jest w znacznym stopniu utrudniony[56]. Tymczasem, środowiska akademickie podkreślają ogromną wagę potrzeby wymiany danych spseudonimizowanych dla celów rozwoju nauki i apelują o pilne rozwiązanie prawne[57]. Jest to szczególnie istotne w przypadku projektów badawczych polegających na stworzeniu wspólnych baz danych, które mają być wykorzystywane przez innych naukowców. Jeśli w bazach będą znajdować się dane osobowe, udostępnianie ich podmiotom spoza EOG nawet w formie spseudonimizowanej powinno się odbywać zgodnie z wymaganiami art. 46 i n. RODO.

Prawa osób, których dane dotyczą

Kolejnym istotnym elementem zgodności, który trzeba wziąć pod uwagę decydując o sposobie wykorzystania danych jest **możliwość i konieczność realizacji żądań osób, których dane dotyczą**.

W przypadku gdy dane są anonimowe, zasady RODO nie obowiązują. Zatem, osobom których dane zostały włączone do danych projektu nie przysługują szczególne prawa względem tych danych.

Projekt korzystający z danych spseudonimizowanych musi rozważyć zgodność z przepisami art. 15–21 RODO, które przyznają uczestnikom badania m.in. prawo dostępu do danych, prawo do sprostowania i usunięcia danych, prawo do ograniczenia przetwarzania danych, prawo do przenoszenia danych, prawo do sprzeciwu oraz – potencjalnie – prawo do wycofania zgody

[52] Przykład takiego planu, https://www.fsd.tuni.fi/en/services/data-management-guidelines/anf-template.pdf.
[53] *M. Hintze, K. El Emam*, Comparing the Benefits…, *op. cit.*, s. 7.
[54] *E. Gourd*, GDPR obstructs cancer research data sharing, „The Lancet Oncology", vol. 22, Nr 5/2021, s. 592.
[55] Wyrok TSUE z 16.7.2020 r. w sprawie C-311/18, *Komisarz ds. ochrony danych* przeciwko *Facebook Ireland Ltd i Maximillian Schrems*, Legalis.
[56] Raport ALLEA, EASAC and FEAM International Sharing of Personal Health Data for Research, https://www.feam.eu/wp-content/uploads/International-Health-Data-Transfer_2021_web.pdf.
[57] *Ibidem*.

na przetwarzanie danych. Od powyższych zasad istnieją jednak ważne wyjątki. Źródło tych wyjątków jest dwojakie. Po pierwsze, mogą mieć zastosowanie ogólne wyjątki przewidziane w RODO. W szczególności:

1) w przypadku pozyskiwania danych z innych źródeł, niż osoba, której dane dotyczą, prawo do informacji może zostać ograniczone, jeżeli jej udzielenie niezbędnych informacji byłoby niemożliwe lub wymagałoby nieproporcjonalnie dużego wysiłku. Ponadto, gdy konsekwencje obowiązku informacyjnego mogą uniemożliwić lub poważnie utrudnić osiągnięcie celów przetwarzania do celów badań naukowych, administrator może być zwolniony z obowiązku informacyjnego (art. 14 ust. 5 lit. b RODO);
2) podobnie, prawo do bycia zapomnianym nie ma zastosowania, jeżeli jego wykonanie prawdopodobnie uniemożliwiłoby lub poważnie utrudniło osiągnięcie celów przetwarzania do celów naukowych (art. 17 ust. 3 RODO);
3) jeśli chodzi o prawo do sprzeciwu i badania naukowe, osoba, której dane dotyczą, ma prawo, z przyczyn związanych z jej szczególną sytuacją, sprzeciwić się przetwarzaniu danych osobowych, chyba że przetwarzanie jest konieczne do wykonania zadania realizowanego ze względu na interes publiczny (art. 21 ust. 6 RODO).

Dodatkowo, zgodnie z art. 12 ust 2 RODO, jeżeli administrator danych może wykazać, że nie jest w stanie zidentyfikować osoby, której dane dotyczą, na podstawie danych, które posiada, nie musi przestrzegać praw z art. 15–21 RODO. W swojej opinii EIOD uznaje również, że w określonych okolicznościach duża liczba osób sprzeciwiających się całości lub części badań naukowych może mieć negatywny wpływ na reprezentatywność i wiarygodność danych badawczych, a tym samym na rzetelność badań, np. w działalności badawczej związanej z rzadkimi chorobami[58].

Po drugie, RODO upoważnia kraje członkowskie do wprowadzenia wyjątków od zasad przewidzianych w art. 15, 16, 18 i 21 RODO w swoich systemach prawnych[59]. Zatem, wyjątki w tym zakresie nie są zharmonizowane i różnią się w zależności od jurysdykcji.

Podsumowanie

Celem niniejszego artykułu było przedstawienie wybranych zagadnień związanych z przetwarzaniem anonimowych lub spseudonimizowanych danych osobowych, mających wpływ na podmioty prowadzące badania naukowe. W dobie digitalizacji oraz szerokiego dostępu do różnych zbiorów informacji, właściwa ochrona tożsamości uczestników badania wymaga ciągłej ewaluacji przyjętej metody de-identyfikacji danych. Kwestia ta wymaga szczegółowej oceny, w zależności od kontekstu projektu badawczego. Głównym czynnikiem ryzyka, który może prowadzić do re-identyfikacji osób fizycznych na podstawie danych uprzednio zanonimizowanych, jest połączenie lub dopasowanie danych z jednego lub kilku innych źródeł z danymi pozornie anonimowymi[60]. Wówczas może okazać się, że dane, z których korzysta lub które upublicznia projekt badawczy nie są anonimowe, a „ucieczka" spod reżimu RODO była przedwczesna. Stąd też należy rozważyć czy silna pseudonimizacja danych nie zapewni większego bezpieczeństwa dla uczestników badania. Podobnie jak anonimizacja chroni ona osoby fizyczne przed ujawnieniem ich tożsamości osobom postronnym. Z drugiej strony, przetwarzanie danych spseudonimizowanych wymaga spełnienia przez projekt badawczy szeregu rygorystycznych wymogów RODO przez cały czas trwania projektu. W konsekwencji, poziom ochrony danych spseudonimizowanych, jak też możliwość kontroli ich przetwarzania przez osoby fizyczne, będą korzystnie wpływały na prawa i wolności osób.

ABSTRACT

Personal data pseudonymization and anonymization in research – selected issues

The aim of the article is to present the issues of pseudonymization and anonymization of personal data in the context of scientific research. The first part of the paper outlines the definitions of the above concepts in light of the provisions of the GDPR and points out practical issues and controversies related to the processes of de-identification of information about individuals to be used in research projects. In its second part, the publication presents some of the consequences of the choice of a data processing method, in particular as regards the legal basis for the use of data, risks for individuals, and the exercise of rights under the GDPR.

Słowa kluczowe: anonimizacja, pseudonimizacja, badania naukowe, dane osobowe
Key words: anonymization, pseudonymization, scientific research, personal data

INCISIVE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952179. However, the content of this article reflects the opinion of its author and does not in any way represent opinions of the European Union. European Commission is not responsible for any use that may be made of the information the article contains.

[58] Wstępna opinia EIOD dotycząca ochrony danych i badań naukowych (EDPS, *A Preliminary Opinion on data protection and scientific research),* 6.1.2020 r., s. 21, https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf.

[59] Artykuł 89 ust. 2 oraz art. 5 ust. 5 lit. b) RODO.

[60] *J. Henriksen-Bulmer, S. Jeary*, Re-identification attacks–A systematic literature review, „International Journal of Information Management", vol. 36, Nr 6/B/2016, s. 1184–1192, https://doi.org/10.1016/j.ijinfomgt.2016.08.002.