

Всероссийская конференция «Биомедицинская химия: наука и практика»

ПРОТОКОЛЫ ЭКСПЕРИМЕНТОВ, ПОЛЕЗНЫЕ МОДЕЛИ, ПРОГРАММЫ И СЕРВИСЫ

БОЛЬШИЕ ДАННЫЕ НАЦИОНАЛЬНЫХ РЕЕСТРОВ ЛЕКАРСТВЕННЫХ СРЕДСТВ

П.И. Савосина*, Д.С. Дружиловский, Д.А. Филимонов, В.В. Поройков

Научно-исследовательский институт биомедицинской химии им. В.Н. Ореховича,
119121, Москва, ул. Погодинская, 10; *e-mail: polina.savosina@ibmc.msk.ru

В настоящее время при разработке и применении методов компьютерного конструирования лекарств широко используются различные базы данных одобренных лекарственных соединений для составления обучающих выборок и валидации полученных моделей. Большинство свободно доступных баз данных содержит сведения об ограниченном количестве препаратов. Это приводит к уменьшению изучаемого и используемого исследователями фармакотерапевтического химического пространства. Сведения о препаратах, разработанных и применяемых локально, в одной-двух странах, могут быть получены из соответствующих национальных реестров лекарственных средств. Нами были найдены и проанализированы реестры более 70 стран мира, доступ к которым осуществляется посредством открытых веб-ресурсов в сети Интернет. Помимо перечней одобренных терапевтических средств, во многих веб-ресурсах представлена возможность изучать официальные документы, публикуемые медицинскими ведомствами после процесса одобрения препарата. Эти документы содержат широкий спектр информации о лекарствах, включая данные о структурных формулах, фармакодинамике и фармакокинетике, токсичности и т.д. Представленные в реестрах соединения как по количеству, так и по структурному разнообразию превосходят информацию, представленную в широко используемых базах данных одобренных препаратов. Агрегированные сведения из национальных реестров лекарств представляют собой пример “больших данных” в биомедицинской области, учитывающий все затруднения, возникающие при их обработке. Использование этой информации в компьютерном конструировании лекарств позволит не только расширить исследуемое фармакотерапевтическое химическое пространство, но и повысить качество получаемых (Q)SAR моделей.

Ключевые слова: “большие данные”; лекарственные средства; национальные реестры лекарственных средств; медицинские агентства; веб-ресурсы; World Wide Approved Drugs

DOI: 10.18097/BMCRM00230

ВВЕДЕНИЕ

В настоящее время информация об одобренных лекарственных соединениях используется в разнообразных биомедицинских исследованиях как для изучения свойств уже известных, так и для разработки новых фармакологических препаратов. В связи с этим в сети Интернет представлено большое количество баз данных (БД), содержащих информацию о таких соединениях [1]. В компьютерном конструировании лекарств информация из подобных ресурсов используется для построения различных обучающих выборок, на основе которых выполняется прогностическая оценка физико-химических свойств или биологических эффектов исследуемых соединений. Количество и качество исходных данных играют решающее значение для эффективного применения методов *in silico*. Поэтому наиболее часто исследователи используют сведения об одобренных лекарственных соединениях, свойства и характеристики которых были детально изучены для получения разрешения на медицинское применение.

Однако в настоящее время не существует единой свободно доступной БД, содержащей информацию обо всех лекарственных средствах, когда-либо одобренных к медицинскому применению [2].

Большая часть широко используемых веб-ресурсов содержит сведения об ограниченном количестве лекарств, в основном только о препаратах, одобренных в США и Европе. Достоверная информация о терапевтических средствах, разработанных и используемых локально, в одной-двух странах, в связи, например, с этнической принадлежностью, остаётся неизвестной [3]. Это приводит к снижению изучаемого и используемого исследователями фармакотерапевтического химического пространства.

Информация о локально используемых лекарствах может быть получена из национальных реестров лекарственных средств (НРЛС). Медицинские агентства многих стран размещают свои НРЛС в сети Интернет для свободного доступа к ним исследователям и специалистам из области здравоохранения. Тем не менее, поиск таких реестров затруднён из-за большого количества регулирующих органов в мире [4], отличий представления информации, а также отсутствия единого ресурса, обеспечивающего доступ к этим данным [2].

Целью нашей работы был поиск свободно доступных НРЛС различных стран и анализ возможностей и ограничений использования представленной в них информации, а также сопоставление найденной информации со сведениями из широко используемых БД одобренных препаратов.



1. СВОБОДНО ДОСТУПНЫЕ НАЦИОНАЛЬНЫЕ РЕЕСТРЫ ЛЕКАРСТВ

Поскольку в настоящее время отсутствует единый информационный ресурс, обеспечивающий доступ к НРЛС, мы провели поиск веб-сайтов медицинских ведомств различных стран в сети Интернет. На основе полученных данных были найдены веб-ресурсы, обеспечивающие свободный доступ к информации о лекарственных препаратах, разрешенных к медицинскому применению в каждой конкретной стране. На текущий момент нами были найдены веб-ресурсы для 71 страны и региона (см. Дополнительные материалы).

2. ОБЪЁМ ПРЕДСТАВЛЕННЫХ В НРЛС ДАННЫХ

Для 49 из найденных 71 НРЛС мы подсчитали количество представленных в них записей. Для остальных реестров это сделать не удалось, по причине ограниченной функциональности соответствующих ресурсов в предоставлении всей имеющейся в них информации. Как видно из рисунка 1, количество представленных записей в проанализированных НРЛС сильно различается. В первую очередь это связано с тем, что в реестрах некоторых стран помимо фармацевтических препаратов представлены также косметические и гигиенические средства, ветеринарные препараты, средства традиционной медицины и гомеопатии. В связи с этим необходимо проводить предварительную фильтрацию данных для получения сведений только о веществах, относящихся к необходимой категории.

Большое число записей связано со способом ведения НРЛС. Так, в реестрах России и Казахстана при каждом изменении как статуса регистрационного удостоверения (одобрение, продление, исключение и т.д.), так и текста официальной документации (новое показание к применению, выявленные в ходе пост-маркетинговых исследований побочные эффекты и т.д.) происходит появление новой записи. С одной стороны, это позволяет отслеживать в динамике результаты клинического использования препарата. С другой — приводит к необходимости обрабатывать большее количество записей с их последующей агрегацией.

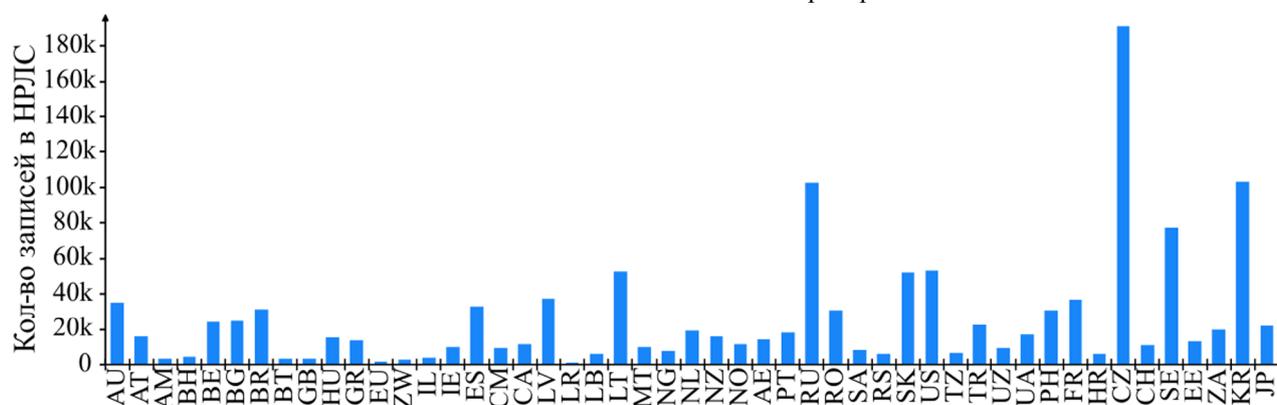


Рисунок 1. Количество записей в 49 найденных национальных реестрах лекарственных средств. Здесь и далее наименования стран приведены в виде двухбуквенных кодов стандарта ISO 3166-1.

3. ИНФОРМАЦИЯ О ЛЕКАРСТВАХ, ДОСТУПНАЯ В НРЛС

Найденные реестры можно условно разделить на две основные группы. Первая группа — краткие табличные реестры, содержащие торговое и непатентуемое наименование, дозировки и способы приёма препаратов. Используя их, можно получить полный перечень одобренных терапевтических средств. Однако, более детальные сведения, например, показания к применению лекарств, предоставлены во второй группе, к которой относятся веб-ресурсы. Используя их, можно получить доступ к следующим официальным документам: Общая характеристика препарата (Summary of Product Characteristics, SmPC), FDA Drug Label, Монография лекарства, Инструкция по применению и Обзорный отчёт регулирующего органа.

Все публикуемые медицинскими ведомствами официальные документы содержат сведения о: (1) наименованиях (международных или национальных) лекарств, (2) вспомогательных веществах, (3) терапевтических назначениях, (4) способах применения, (5) дозировках, (6) побочных эффектах, (7) противопоказаниях и т.д. (рис. 2). Помимо этого, во всех типах документов представлена информация об известных механизмах действия лекарств (включая наименования молекулярных мишеней и типы взаимодействия с ними), токсичности, фармакокинетики (в том числе сведения о метаболизирующих ферментах и взаимодействии с транспортерами) и межлекарственных взаимодействиях. Однако для многих препаратов механизм действия либо достоверно не установлен (известны только виды фармакологических эффектов, которые возникают при приёме препарата), либо взаимосвязь молекулярной мишени с терапевтическим эффектом не известна. Информация о фармакокинетики лекарств, применяемых наружно, также представлена крайне редко.

Обзорные отчёты регулирующих органов, FDA Drug Label и монографии лекарств содержат в том числе данные о структурных формулах лекарственных соединений и их физико-химических характеристиках. Также обзорные отчёты содержат подробную информацию о результатах исследований вторичной фармакодинамики и клинических испытаний препаратов.

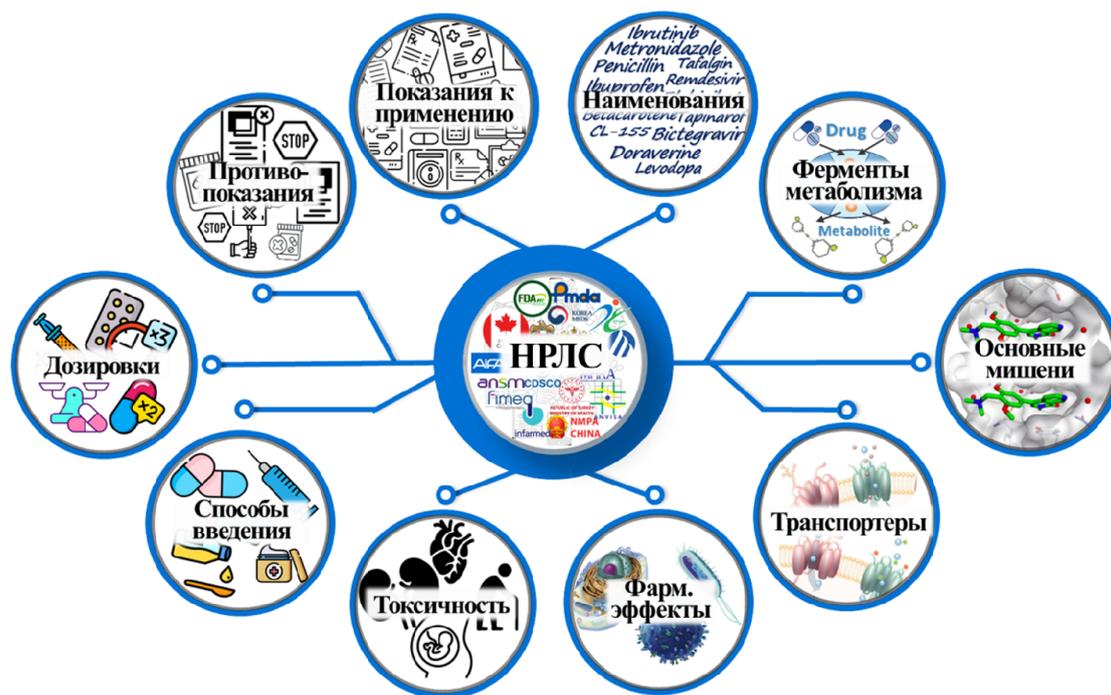


Рисунок 2. Сведения о лекарственных препаратах, доступные во всех типах официальных документов, публикуемых медицинскими регулирующими ведомствами после процесса одобрения. NMRs – национальные реестры лекарственных средств.

Дополнительным преимуществом является наличие в некоторых реестрах, например НРЛС России, сведений о производителях фармацевтической субстанции. Эта информация может быть использована, в частности, для поиска поставщиков образцов лекарственных соединений.

4. КОЛИЧЕСТВО ОДОБРЕННЫХ ЛЕКАРСТВ ПО ДАННЫМ НРЛС

Для 49 из исследованных реестров мы подсчитали количество представленных в них уникальных низкомолекулярных активных фармацевтических субстанций (рис. 3А). Среднее количество таких соединений, представленных в НРЛС по данным на 2024 год, составляет 1057 веществ. В большинстве стран, веб-ресурсы медицинских ведомств которых были проанализированы, разрешены к применению “локальные” препараты, не встречающиеся в других реестрах. Как видно из рисунка 3Б, сведения о менее чем 50 низкомолекулярных активных фармацевтических субстанций содержатся во всех проанализированных реестрах. При этом, более 1100 лекарственных соединений присутствуют в НРЛС не более двух стран.

5. ОСОБЕННОСТИ ОБРАБОТКИ ДАННЫХ НРЛС

Основную сложность в обработке данных НРЛС создают различия в: (1) методах доступа к данным, (2) способах представления реестра, (3) типах публикуемых документов и (4) представлении данных на государственном языке. Важно также отметить, что в некоторых НРЛС, например, Американском, Российском, Белорусском, в качестве названия

лекарственного соединения указывается наименование активного метаболита, а не той субстанции, которая изначально присутствует в препарате. Наименование субстанции можно получить только из официальной документации. Это влечёт за собой трудности в сопоставлении данных из различных реестров.

Также дополнительную сложность создаёт представление текста документов в виде изображений, как, например, в реестрах России и США. На сегодняшний день доступные для использования алгоритмы интеллектуального анализа текстов и распознавания изображений (в том числе и структурных формул соединений) [5, 6] не позволяют с достаточно высокой точностью обработать доступную информацию. Их использование для обработки массива официальных документов пока невозможно.

6. БАЗА ДАННЫХ ФАРМАЦЕВТИЧЕСКИХ СУБСТАНЦИЙ WWAD

Основываясь на полученной из медицинских реестров информации об одобренных к клиническому применению лекарственных средствах, нами была создана база данных World Wide Approved Drugs (WWAD) [7]. Текущая версия нашей БД содержит информацию о более чем 4400 уникальных низкомолекулярных фармацевтических субстанциях, в том числе сведения об их структурной формуле, взаимодействии с терапевтическими и вторичными мишенями, фармакологических эффектах, терапевтических назначениях и разрешении к медицинскому применению в различных странах. Нами было получено Свидетельство о государственной регистрации базы данных “World Wide Approved Drugs” № 2024621221 от 21 марта 2024 г.

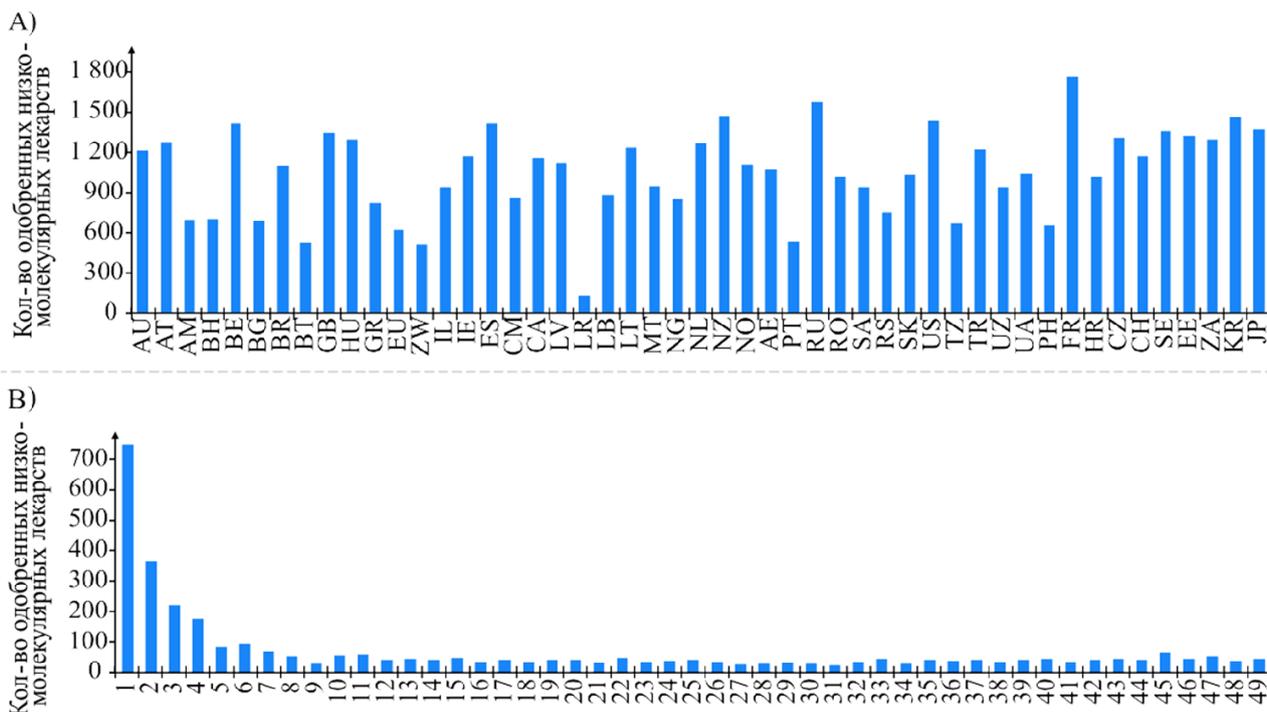


Рисунок 3. Количество уникальных низкомолекулярных органических лекарственных соединений: А) представленных в 49 проанализированных НРЛС; В) в зависимости от количества НРЛС, в которых они присутствуют.

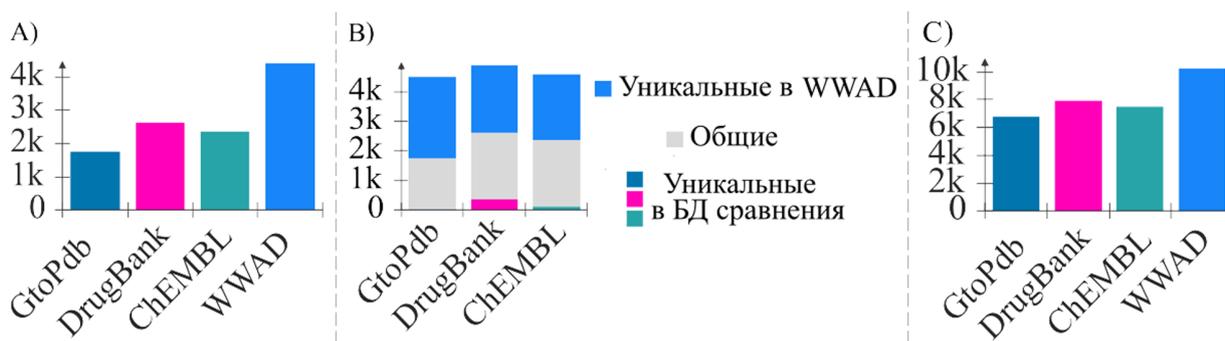


Рисунок 4. Анализ структур уникальных низкомолекулярных лекарственных соединений, представленных в WWAD, DrugBank, ChEMBL и GtoPdb. А) Количество структур соединений, представленных в БД; В) количество уникальных и общих структур соединений между WWAD и БД сравнения; С) количество уникальных MNA дескрипторов для всех соединений сравниваемых БД.

Мы оценили пересечение и структурное разнообразие лекарственных соединений, собранных из 71 НРЛС и представленных в WWAD и трёх широко используемых веб-ресурсах: DrugBank [8], ChEMBL [9] и IUPHAR/BPS Guide to PHARMACOLOGY (GtoPdb) [10]. Для сравнительного анализа из этих трёх веб-ресурсов были экспортированы файлы в формате SDF, содержащие структуры и наименования соединений, которые отмечены в каждом из них как “одобренные лекарства”.

Для оценки структурного разнообразия мы использовали разработанные нами ранее MNA дескрипторы [11]. Данный способ описания структурной формулы представляет собой список уникальных строк, отображающих её фрагменты. Подсчёт количества уникальных MNA дескрипторов для всех молекул в каждой БД предоставляет

возможность оценить наличие в ней структурно различных соединений.

По результатам сравнения было выявлено, что количество лекарственных соединений в WWAD значительно превосходит число молекул соответствующей категории в базах данных сравнения (рис. 4А). При этом число уникальных и структурно различных веществ в созданной нами БД также значительно выше (рис. 4В и 4С).

ЗАКЛЮЧЕНИЕ

Отсутствие единого ресурса, предоставляющего доступ к НРЛС всех стран, создает определённые трудности при попытках анализа используемых в мире лекарственных субстанций с целью получения доступной информации об их биохимических,

биофизических и фармакотерапевтических особенностях. Дополнительную сложность создаёт отсутствие доступа к некоторым НРЛС для международных исследователей, а также представление всей информации на различных языках.

К настоящему моменту нам удалось найти свободно доступные реестры лекарств из 71 регионального и национального медицинского агентства. Анализ большого количества НРЛС позволил выявить терапевтические средства, разработанные и используемые вне США и Европы. Официальные документы, публикуемые медицинскими ведомствами после процесса одобрения препарата, содержат широкий спектр информации о лекарственном соединении, которая может быть использована в методах компьютерного конструирования лекарств. Эта информация позволила создать базу данных WWAD, по сравнению с которой другие свободно доступные ресурсы существенно отстают в отражении современного состояния фармацевтического представления и структурного разнообразия низкомолекулярных лекарственных соединений в мире.

В процессе анализа собранных в БД WWAD массива гетерогенных биомедицинских и химических данных, выяснилось, что несмотря на сравнительно малое количество представленных на фармацевтических рынках мира лекарственных низкомолекулярных фармацевтических субстанций, объём и разнообразие информации о них является примером “больших данных” в биомедицинской области. Учтя, что для сбора, последующей обработки и обогащения таких данных необходима разработка специальных алгоритмов их анализа, стандартизация и нормализация всеобъемлющих данных об известных лекарствах позволят использовать в исследовательских целях наиболее достоверную информацию о лекарственных препаратах и расширить исследуемое известное фармакотерапевтическое химическое пространство.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Работа не связана с исследованиями, в которых в качестве объекта выступают люди или животные.

БЛАГОДАРНОСТИ

Авторы выражают искреннюю признательность компании ChemAxon за предоставление лицензии на программный комплекс Instant JChem.

ФИНАНСИРОВАНИЕ

Работа выполнена в рамках Программы фундаментальных научных исследований в Российской Федерации на долгосрочный период (2021–2030 годы) (№ 122030100170-5).

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

К данной статье приложены дополнительные материалы, свободно доступные на сайте журнала (<http://dx.doi.org/10.18097/bmcrm00000>).

ЛИТЕРАТУРА

1. Tanoli, Z., Seemab, U., Scherer, A., Wennerberg, K., Tang, J., Vähä-Koskela, M. (2021) Exploration of databases and methods supporting drug repurposing: A comprehensive survey. *Briefings Bioinformatics*, **22**(2), 1656–1678. DOI: 10.1093/bib/bbaa003
2. Siramshetty, V.B., Grishagin, I., Nguyễn, Đ.T., Peryea, T., Skovpen, Y., Stroganov, O., Katzel, D., Sheils, T., Jadhav, A., Mathé, E.A., Southall, N.T. (2022) NCATS Inxight Drugs: A comprehensive and curated portal for translational research. *Nucleic Acids Res.*, **50**(D1), D1307–D1316. DOI: 10.1093/nar/gkab918
3. Maglo, K.N., Mersha, T.B., Martin, L.J. (2016) Population genomics and the statistical values of race: An interdisciplinary perspective on the biological classification of human populations and implications for clinical genetic epidemiological research. *Front. Genet.*, **7**, 22. DOI: 10.3389/fgene.2016.00022
4. World Health Organization. Competent authorities of countries participating in the WHO. Retrieved June 11, 2024, from: <https://www.who.int/teams/regulation-prequalification/regulation-and-safety/regulatory-convergence-networks/certification-scheme/contacts>
5. Leaman, R., Islamaj, R., Adams, V., Alliheedi, M.A., Almeida, J.R., Antunes, R., Bevan, R., Chang, Y.C., Erdengasileng, A., Hodgskiss, M., Ida, R., Kim, H., Li, K., Mercer, R.E., Mertová, L., Mobasher, G., Shin, H.C., Sung, M., Tsujimura, T., Yeh, W.C., Lu, Z. (2023) Chemical identification and indexing in full-text articles: An overview of the NLM-Chem track at BioCreative VII. *Database*, **2023**, baad005. DOI: 10.1093/database/baad005
6. Chaudhuri, A., Mandaviya, K., Badelia, P., Ghosh, S.K. (2017) Summary and Future Research. In: *Optical Character Recognition Systems for Different Languages with Soft Computing*. Studies in Fuzziness and Soft Computing (Chaudhuri A., Mandaviya K., Badelia P., Ghosh S.K., eds.) Springer: Cham, Switzerland, pp. 241–245.
7. World Wide Approved Drugs database. Retrieved June 11, 2024, from: <https://www.way2drug.com/wwad>
8. DrugBank database. Approved drugs. Retrieved June 11, 2024, from: <https://go.drugbank.com/drugs>
9. ChEMBL database. Approved drugs. Retrieved June 11, 2024, from: <https://www.ebi.ac.uk/chembl>
10. IUPHAR/BPS Guide to PHARMACOLOGY database. Approved drugs. Retrieved June 11, 2024, from: <https://www.guidetopharmacology.org/GRAC/LigandListForward?type=Approved>.
11. Filimonov, D., Druzhilovskiy, D., Lagunin, A., Glorizova, T., Rudik, A., Dmitriev, A., Pogodin, P., Poroikov, V. (2018) Computer-aided prediction of biological activity spectra for chemical compounds: Opportunities and limitations. *Biomedical Chemistry: Research and Methods*, **1**(1), e00004. DOI: 10.18097/BMCRM00004

Поступила: 15. 06. 2024.

После доработки 28. 06. 2024.

Принята к публикации: 01. 07. 2024.

BIGDATA OF NATIONAL MEDICINE REGISTERS*P.I. Savosina*, D.S. Druzhilovskiy, D.A. Filimonov, V.V. Poroikov*Institute of Biomedical Chemistry,
10 Pogodinskaya str., Moscow, 119121 Russia; *e-mail: polina.savosina@ibmc.msk.ru

Currently, various databases of approved drugs are widely used in the development and application of computer-aided drug design for training sets and validating predicted models. Most of the freely available databases contain information on a limited number of drugs. This leads to reduction in the space of pharmacotherapeutic chemistry studied and used by researchers. Information on drugs developed and used locally, in one or two countries, can be obtained from relevant national medicine registries. We have identified and analyzed registries from more than 70 countries around the world that are accessible through open web resources on the Internet. In addition to lists of approved therapeutics, many web resources offer the opportunity to review official documents published by medical authorities after the approval process. These documents contain a wide range of information about drugs, including data on structural formulations, pharmacodynamics and pharmacokinetics, toxicity, etc. The compounds represented in the registries, both in terms of quantity and structural diversity, exceed the known widely used databases of approved drugs. Combined data from national drug registries represent an example of “Big Data” in the biomedical field, taking into account all the difficulties involved in their processing. Its use in computer-aided drug design will not only expand the pharmacotherapeutic chemical space studied, but also improve the quality of the original data.

Key words: BigData; approved drugs; national medicine registries; medicine agencies; web-resources; World Wide Approved Drugs

FUNDING

The work was performed within the framework of the Program for Basic Research in the Russian Federation for a long-term period (2021–2030) (No. 122030100170-5).

Received: 15.06.2024; revised: 28.06.2024; accepted: 01.07.2024.