KI - Künstliche Intelligenz
https://doi.org/10.1007/s13218-023-00809-6

DISCUSSION

Opportunities and Challenges for AI-Based Analysis of RWD in Pharmaceutical R&D: A Practical Perspective

Merle Behr[1] · **Rolf Burghaus**[2] · **Christian Diedrich**[2] · **Jörg Lippert**[2]

Received: 14 August 2023 / Accepted: 2 September 2023

Abstract
Real world data (RWD) has become an important tool in pharmaceutical research and development. Generated every time patients interact with the healthcare system when diagnoses are developed and medical interventions are selected, RWD are massive and in many regards typical big data. The use of artificial intelligence (AI) to analyze RWD seems an obvious choice. It promises new insights into medical need, drivers of diseases, and new opportunities for pharmacological interventions. When put into practice RWD analyses are challenging. The distributed generation of data, under sub-optimally standardized conditions in a patient-oriented but not information maximizing healthcare transaction, leads to a high level of sparseness and uncontrolled biases. We discuss why this needs to be addressed independent of the type of analysis approach. While classical statistical analysis and modeling approaches provide a rigorous framework for the handling of bias and sparseness, AI methods are not necessarily suited when applied naively. Special precautions need to be taken from choice of method until interpretation of results to prevent potentially harmful fallacies. The conscious use of prior medical subject matter expertise may also be required. Based on typical application examples we illustrate challenges and methodological considerations.

Keywords Real world data · Artificial intelligence · Machine learning · Pharmaceutical research

1 Introduction

Pharmaceutical innovation differs in many regards from the development of other products. A pharmacological intervention impacts on patient lives in a very elementary way. It is expected to provide a medical benefit that can extend life, prevent morbidity, or at least directly improve quality of life. If something goes wrong, drugs can cause harm. Unintended adverse events can deteriorate quality of life. In severe cases drugs can cause death of patients.

This extreme ambivalence of the potential effects of a pharmaceutical product translates into a unique level of ethical requirements and is reflected in a high level of legal regulation not only for the products but also of the Research and Development (R&D) process itself. R&D in the pharmaceutical industry is regulated both at the level of legislation of individual countries and globally via the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). The ICH Guideline for Good Clinical Practice,[1] only one guidance in a universe of binding recommendations for pharmaceutical industry, provides an impression of the complexity and constraints of drug development.

Unsurmountable ethical and legal constraints but also operational and economic considerations prevent most of the scientifically relevant experiments with animals and humans in pharmaceutical R&D. Altogether this leads to a poor productivity of pharmaceutical R&D. More than a decade from ideation to product launch and more than a billion Euros capitalized R&D investment per new product slow down medical innovation and has been a matter of debate in the pharmaceutical community for decades [18].

A very specific challenge of pharmaceutical researchers is the limited understanding of 'the patient'. An investigation of questions like 'what is the medical need?', 'how does disease progress?', 'how do patients respond to existing Standard of Care (SoC)?', 'how do risks for poor outcome

✉ Merle Behr
merle.behr@ur.de

[1] Faculty of Informatics and Data Science, University of Regensburg, Bajuwarenstraße 4, 93053 Regensburg, Germany

[2] Pharmacometrics, Pharmaceuticals R&D, Bayer AG, Leverkusen, Germany

[1] https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf.

Published online: 09 October 2023

or positive response to treatment manifest in early changes of medical parameters?' are subject to all the limitations discussed above.

As in other research fields, AI has also generated a lot of interest by pharmaceutical researchers. The idea that these questions could be addressed by learning from data in a broader sense than by hypothesis driven experimentation and statistical evaluation of experimental results is compelling. Consequently, in the absence of the possibility to adequately address all research questions with dedicated experiments, i.e. clinical trials, the community is aiming for the secondary use of pre-existing data together with AI as a potential way out. This secondary use does of course include data from clinical studies undertaken for other reasons, in most of the cases for the demonstration of the efficacy and safety of a specific single drug. In addition, today, data generated in daily clinical and non-clinical routine is also considered of high value. The term Real World Data (RWD) has been coined for information collected from electronic health records in outpatient practices and clinics, from claims data used for reimbursement of medical services and drug prescriptions, and from wearable medical devices and health apps used in 'daily life' of patients.[2] In contrast to clinical study data such information is not the result of an experiment with a Design of Experiment dedicated to the investigation of a prespecified scientific question but originates from an uncontrolled observational setting. It suffers from diverse problems. Quality of individual data items is limited by time pressure in healthcare providing institutions and the fact that millions of individual healthcare professionals are contributing to its generation. The sampling time point and even the question if data generated and documented at all depends on highly personal decisions of the patient and its caring physician. RWD has lower quality than clinical trial data in this regard. It is sparse and incomplete. In exchange, these deficits are countered by the breadth and the enormous amounts of data that are—at least in theory—available in our healthcare systems and the fact that the cost of RWD results from data aggregation and curation only while their mere existence comes for free.

The high interest in RWD has already led to regulatory recognition[3] and it is driving the rapid growth of the number of available databases and the research field. The diversity is high. Data can represent large general populations such as the Explorys database[4] covering all structured medical information for more than 50 million US citizens over a period of more than two decades. Other databases have a narrow disease indication focus such as the European sleep apnea cohort ESADA[5] or focus on the point of data generation such as the MIMIC (IV) database[6] where all data originates from the intensive care units (ICU) of a Boston based network of clinics (for ESADA it is a European network of sleep labs) and cover in the order of tens of thousands of patients and short interactions with the healthcare systems (single night stays up to several days to few weeks in the ICU). RWD can be structured and coded according to standardized medical terminology such as ICD, ATC, SNOMED, or LOINC codes or it may include unstructured information such as physician notes (e.g., MIMIC IV only). Some databases contain raw data from medical devices and imaging and diagnostics devices (e.g., MIMIC IV). For more details on different types of RWD see e.g., [13, 14].

RWD is often referred to as a type of Big Data in the medical domain. The dimensionality of RWD can be very high. For example, the Explorys database with more than 50 million patients lists more than 20.000 distinct lab measurements, thousands of medical conditions and interventions. The underlying 'true' dimensionality of 'the patient' is actually not known but may be much higher. A simple example may illustrate this. RWD typically does not contain genomic information about our 20.000 genes and the 300 million polymorphisms (i.e. genetic variants) identified so far. The dimensionality of the available data is very likely only the tip of the iceberg but in comparison to the apparently Big Data like number of patients in RWD sources it puts a question mark to this classification. For many potentially interesting research questions the relevant size category of the data set will be 'small'. We will further discuss this problem below.

Many other specifics of RWD could be discussed such as the difference between care information versus insurance data, the dominance of data from regions with more liberal data privacy legislations and more advanced digitalization of the healthcare sector, and the relevance of features of the coding systems. We will focus on conceptual and methodological challenges for the application of AI to RWD, to ground the co-occuring hype of RWD and AI in the realities of practical applications for pharmaceutical R&D.

2 Typical Questions Addressed with RWD

In order to get more specific about value cases for extracting information from RWD, we will provide several illustrative examples of tasks and questions which are typically addressed by RWD analyses in pharmaceutical R&D. This choice of examples is influenced by our personal experience

[2] https://en.wikipedia.org/wiki/Real_world_data.

[3] https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence.

[4] https://www.ibm.com/downloads/cas/NNPN9J9Q.

[5] https://esada.med.gu.se/.

[6] https://physionet.org/content/mimiciv/2.2/.

in pharmaceutical R&D. We do not intend to be comprehensive but rather discuss typical questions. One possible way of grouping questions that are addressed with RWD analyses is along the time axis where they are most relevant in pharmaceutical research, starting with basic pharmaceutical research questions, which are typically at the beginning of any drug development pipeline, continuing with specific pharmaceutical development questions, which are typically related to specific clinical trial activities, and ending with questions that finally arise in clinical practice.

2.1 Pharmaceutical Research

At the beginning of any pharmaceutical research one first needs to gain a general understanding of a patient population, its medical need and related SoC, as well as risk factors and drivers for a disease. RWD is clearly predestined for these types of questions, as it provides direct insight into the current and past clinical practice and health status of real patients. For example, concrete questions in this context can be as follows:

- Given that a patient has a certain disease or condition (e.g., diabetes), what is the risk that this patient will have a life threatening outcome from this condition (e.g., a major cardiovascular event)?
- Given that a patient has a certain disease or condition (e.g., diabetes), what are typical treatment strategies in real world clinical practice?
- Can a certain patient population be divided into any meaningful groups which differ w.r.t. the specifics of their disease (or care regimen)?

2.2 Pharmaceutical Development

There are several questions that are addressed with RWD which are directly related to classical pharmaceutical development pipelines, especially in the context of randomized clinical trials (RCTs). RCTs are still widely regarded as the 'gold standard' of pharmaceutical R&D and required by health authorities as a basis of market authorizations in most cases. In the planning phase of a prospective RCT, especially when there are no results from other prior comparable RCTs available, e.g., for Proof-of-concept (PoC) trials, there are several concrete questions which are often addressed with RWD, for example:

- What is the expected magnitude for the RCT's endpoint in the control/placebo arm?
- Is a certain biomarker predictive for the RCT's endpoint? The answer to this question might help to find surrogate endpoints for a shorter or leaner PoC trial, e.g., because the surrogate endpoint can be reached faster or can be powered with fewer patients than the actual endpoint of interest.

In situations where a studied drug or a related substance is already on the market (potentially with off-label use in the patient population of interest), RWD can be used to estimate the expected outcome of a RCT, using the 'target trial' concept, see e.g., [11]. This means that one is addressing the following question.

- Given a fully specified study protocol of a hypothetical RCT, what will be the outcome (e.g., observed average treatment effect) of this RCT? Even if the considered target trial is not exactly the same as the planned RCT, the answer to this question can provide valuable insights.

Another important, arguably most direct, application of RWD in the context of RCTs are virtual control arms (VCA). This refers to clinical trials where the RCT is replaced by a treatment-only clinical trial and the control arm is simulated from already available data (e.g., historical RCT data but also RWD). A reason for virtual control arms can be that an actual control arm might be unethical, (e.g., in oncology trials with no available SoC[7]). See [1, 22, 23] for several VCA examples and reasons why VCAs might be preferred over RCTs.

2.3 Clinical Practice

Once a drug enters or is about to enter the market, there are still several questions which might be addressed with RWD. Typical examples are questions regarding pharmacovigilance, such as:

- What are potential rare side-effects which had not been observed previously in RCTs?
- What is the prevalence of side-effects which were already seen in RCTs?

Another area of application is related to repurposing or label-extension of a drug:

- Can a drug which is already on the market for a certain indication also be helpful for another application? See e.g., [26] for an example where RWD was used to address this question.

[7] See e.g., https://clinicaltrials.gov/ct2/show/NCT03737123 for a currently running phase 2 trial in oncology where RWD is used in a virtual control arm.

- Can a drug which is already on the market for a certain population also be helpful for another population (e.g., in pediatrics)?

Finally, there are several applications of RWD in digital health, where the final product is not a drug but rather a software or algorithm, some exemplary questions are:

- How can an algorithm automatically predict from currently available parameters (e.g., blood values) of an ICU patient whether the patient is at high risk (e.g., for acute kidney failure)?
- How can an algorithm automatically classify imaging samples from cancer patients?
- How can an algorithm (e.g., within a smartphone app) recommend certain dietary restrictions to diabetes patients?

In summary, there is a wide spectrum of questions and problems in pharmaceutical research, which might be addressable with RWD. As a general trend we notice that often the characterisation of some underlying mechanism or causal relationship is more the focus of these questions than pure prediction problems. Even in situations where prediction is the primary goal, when the health and well-being of a human life is at stake it is typically required to have insights about the reason why a certain prediction is made. Therefore, standards for quality control, uncertainty quantification, and interpretability of results are much higher than in other data domains. As we will discuss in the following, this has implications for the potential use of AI methods, which are often of black-box nature with little well-understood statistical guarantees.

3 Challenging Features of Real World Data

After having discussed typical questions and problems which might potentially be addressable with RWD analyses, we provide more insights into what RWD typically looks like, what are its challenging and beneficial features. As will be discussed in the next section, these are tightly linked to the respective challenges and benefits of AI methods for RWD. We provide several examples from a large EHR data source, the IBM Watson Health Explorys database (freeze date: January 2021),[8] which covers records of more than 50 million US patients.

[8] The data was supplied by International Business Machines Corporation as part of IBM Explorys Therapeutic Datasets Delivered. Any analysis, interpretation, or conclusion based on these data is solely that of the authors and not International Business Machines Corporation.

3.1 From Large Data to Small Data

The first apparent beneficial feature of RWD is its sheer size, especially compared to sample sizes which are reached in RCTs. While a large phase 3 RCT will typically have less than 10.000 participants, many RWD databases (e.g., IBM Watson Health Explorys) cover a population which exceeds those of an RCT by a factor of several thousands. Moreover, while a typical RCT will only document a hand-full of covariates for each patient, due to cost-efficiency and limitations of the informed consent, in large EHR databases thousands of lab parameters and biomarkers are in principle available. However, because these data are recorded in an event triggered fashion, they are structurally very different to RCT data, i.e. they do not follow a pre-specified protocol and hence feature a very irregular coverage of information. At the same time, obviously, the number of patient records that are available for a particular analysis depends on the indication of interest and its respective prevalence. For these reasons, seemingly huge general purpose data sets very often shrink rather substantially as soon as a more specific question or patient populations are considered. This is demonstrated based on the two prototypical indications: Chronic kidney disease and Primary sclerosing cholangitis.

Example (Chronic Kidney Disease with Macroalbuminuria)

Chronic kidney disease (CKD) is a prevalent condition that is associated with a gradual loss of kidney function over the course of years to decades. According to the KDIGO (Kidney Disease: Improving Global Outcomes) criteria [21] CKD is classified in terms of the estimated glomerular filtration rate (eGFR) and the urine albumin creatinine ratio (UACR). Let us assume that we require an analysis data set with patients with CKD stage 2 or lower (e.g. $< 60\ ml/min/1.73m^2$ eGFR) and so called macroalbimunuria ($UACR \geq 300\ mg/g$). This is a fairly typical setup for an analysis in this indication.

Using the IBM Explorys delivered EHR database we start with a total of 60 Mio patients in the database - 40 Mio of which have lab data available. Out of those 40 Mio about 19 Mio have at least one creatinine measurement recorded so that an estimated GFR may be calculated. So far, we have only 'lost' about 2 thirds of the database. If, however, we now require patients to have eGFR as well as UACR available this already narrows down the number of available subjects to about 1 Mio even before having applied any explicit disease specific criteria. In contrast to creatinine UACR is a non-standard measurement that will usually only be conducted if risk factors for albuminuria such as diabetes mellitus are present. Therefore, compared to a general population there will be a much larger number of patients with CKD in the 1 Mio patients that we are now left with. This implicit 'enrichment' is a result of the

Table 1 Breakdown of sample size for Chronic kidney disease with Macroalbuminuria cohort

Step	Number of patients
Total data base	~ 60 Mio
At least one creatinine measurement	~19 Mio
+ At least one eGFR and one UACR measurement	~1 Mio
+ eGFR < 60	319,053
+ Historic UACR available at same time	174,369
+ UACR ≥ 300 mg/g	24,494

Table 2 Breakdown of sample size for Primary sclerosing cholangitis (PSC) cohort

Step	Number of patients
Total data base	~60 Mio
Cholangitis diagnosis	28,813
+ Any observation	22,533
+ MRCP or ERCP procedure	5092

selection bias that comes with enforcing the availability of UACR. If we now require patients to have an eGFR < 60 and $UACR \geq 300\ mg/g$ we end up with ~ 25 thousand patients (see Table 1). While this still is a large number that will in many cases be perfectly sufficient for meaningful analyses, it still only is the same order of magnitude as the size of late stage clinical trials in this indication.

Example (Primary Sclerosing Cholangitis)

In contrast to CKD, primary sclerosing cholangitis (PSC) is a rare disease, with prevalence in the United States estimated to be between 1 and 16 in 100.000 per year, see [9, 12]. Again we use the IBM Explorys delivered data set in order to create a PSC patient cohort. Firstly, we note that the specific ICD10 code for PSC (K83.01) is not occupied in the database so that we have to use more generic cholangitis codes. This will e.g. include acute forms of cholangitis and will therefore not be sufficiently specific. Using this ICD code filter as a starting point in the cohort definition it will therefore eventually be necessary to further stratify based on liver function markers such as bilirubin (particularly direct bilirubin), ALT and AST. Furthermore, either endoscopic or MRI based specific diagnostics is required in order to diagnose PSC. The cohort definition should therefore include at least the presence of a respective procedure code (Magnetic resonance cholangiopancreatography (MRCP) or endoscopic cholangiopancreatography (ERCD)).

As can be seen from Table 2 there are ~ 29 thousand patients in the database with an ICD code for cholangitis but only 22.5 thousand of these patients have at least one observation recorded in the database. Out of these patients only about five thousand have a procedure code for ERCD or MRCP documented. Considering that about 122.6 Mio patients years are covered by the data, this equates to an annual prevalence of roughly 4 patients per 100.000 patients. This is well within the range known from literature and hence in this case the comparatively low number that we are ending up with is not surprising. It still demonstrates that when conducting an analysis in a rare indication the overall size of a general purpose EHR database naturally shrinks substantially. The final number in Table 2 does not take into account the necessity for further characterisation of the patient's disease state based on liver function markers that will be given in many use cases. If e.g. characterisation in terms of direct bilirubin is needed, which is only available for ~ 38% of the patients, this will further reduce the cohort size accordingly.

In summary, these examples show that even though RWD appears to be of big nature at first glance, it often shrinks down to not that big data for specific questions. On the other hand, it should also be acknowledged that recruitment for a RCT, especially for a rare disease such as PSC, is also very difficult and costly, such that the resulting patient size in the previous example is still likely to exceed the sample size of any hypothetical RCT for PSC.

3.2 Different Mechanisms for Data Generation: Irregular Sampling and Data Coverage

Besides the size of a patient cohort which is eligible for a RWD analysis, there are several implications on the design of real world data based analyses that stem from the different rationale behind data collection. In a clinical trial the longitudinal coverage in terms of type of parameter and sampling scheme is pre-defined in the study protocol. In contrast to that, real world data recording is purely event driven. The triggering events are very diverse and range from routine check-up visits to admission to intensive care units for treatment of life- threatening acute conditions. Depending on the trigger the type of parameters that are being recorded will typically be very different. Consequently, real world data are sparse with respect to availability of parameters. A special biomarker that is only relevant in a particular indication will, for example, only be measured in patients who at least were suspected to be affected by this indication. Moreover, the longitudinal coverage will be varying with very irregular sampling. We illustrate this again with the Explorys PSC cohort from above.

Example (PSC Continued, Missingness, Coverage, Irregular Time Grid)

The impact of sparsity of RWD in terms of availability of parameters has already been demonstrated on the CKD example given above. This is of course even more of an issue in case of rare diseases where the cohort size is anyway

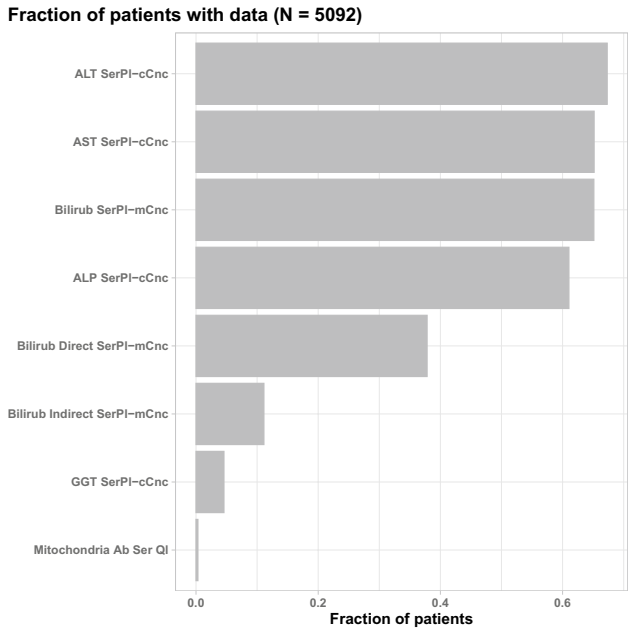

Fig. 1 Fraction of population for which liver markers are available within baseline period from 6 month prior to index event; i.e. first cholangitis diagnosis

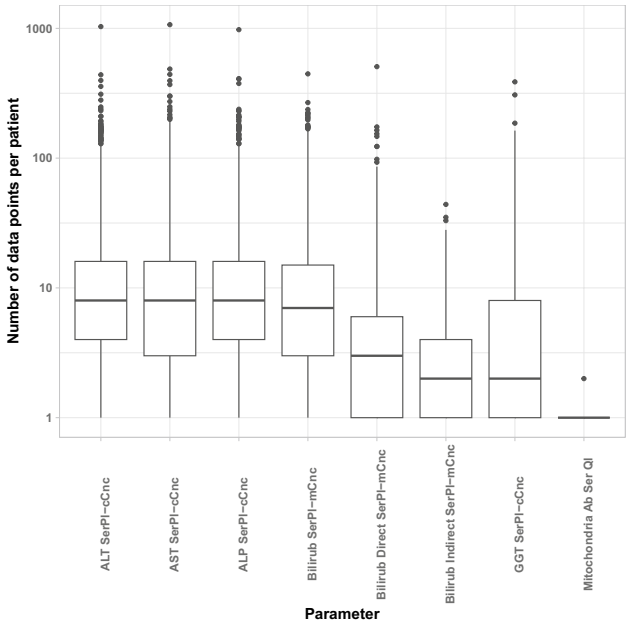

Fig. 2 Distribution of number of data points after the index event

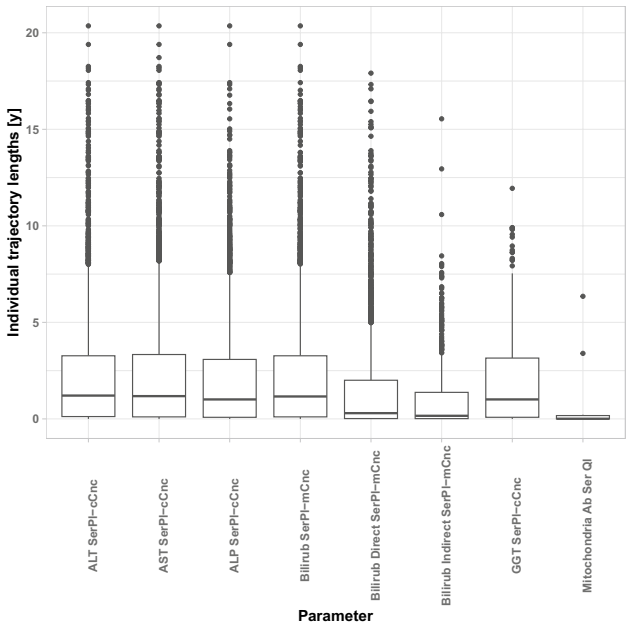

Fig. 3 Distribution of length of follow-up per patient (i.e. data base life after index event)

much smaller. In Fig. 1 this is demonstrated based on several liver markers used for characterizing PSC patients. More general markers such as ALT, AST, bilirubin and also ALP which will in many cases be part of a standard blood test are available for about 60% of the cohort at baseline. Restricting the data set to only patients that have any of these available, will leave us with a data set of 3000 patients which should be sufficient for many relevant questions. Direct bilirubin, which is a more specific marker for hepatic clearance activity than total bilirubin, already is only available for about a third of the population. Restricting ourselves to a population with Gamma-glutamyl Transferase (GGT) information at baseline will further reduce the cohort to only 233 patients. This will probably mean that either imputation is necessary or the data set will simply be too small in many cases. Finally, it is immediately obvious that using mitochondrial antibodies measurements for any analysis data set will be prohibitive in almost all cases as these are only available for 16 patients at baseline. In any case, going with a smaller subset of patient cohorts because of data availability always poses the risk of implicitly enriching a particular type of patient due to selection bias. The consequences of this kind of bias are diverse and impossible to foresee at a general level. They always need to be considered for the case at hand. So far, we have only looked into availability of baseline information. For any kind of progression analysis on the liver markers introduced above it is crucial to have sufficient follow-up data in terms of trajectory length as well as number of samples in follow-up period.

As evident from Figs. 2 and 3 the longitudinal coverage for all parameters will not be satisfactory for many analyses even for parameters like AST and ALT which have a decent availability at baseline. The median follow-up times are only 1–2 years which is short compared to the rate of progression in PSC. The number of available samples on the other hand is comparatively large. For example, patients have a median

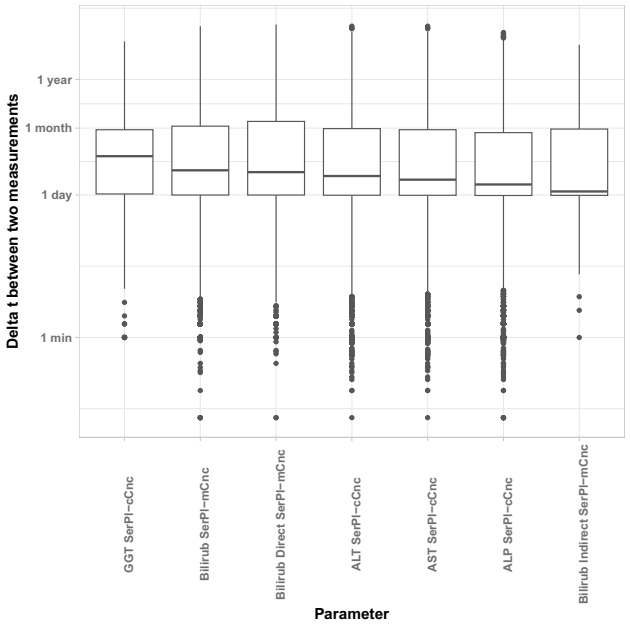

Fig. 4 Distribution of time differences (Δt) between two consecutive measurements of the same kind

number of eight data points available for AST, ALT and ALP. The above mentioned irregularity of sampling times that comes with the event driven data collection poses another crucial challenge for many kinds of longitudinal modeling. Established methods for time series analysis almost always require a regular time grid and hence are not straightforwardly applicable to real world data.

Figure 4 shows this on the basis of the liver markers in the PSC cohort. The inter-quartile range of the Δt distribution (Δt = the time difference between two consecutive measurements) spans from 1 days to 1 month and the tails of the distribution range down to one minute and up to well above one year. Obviously whatever kind of patient trajectory model is used this will have to be dealt with. Typically there will be no way around an explicit handling of time. This is true for conventional methods as well as AI/ML approaches [16].

3.3 Biases and Confounding

Another important aspect related to the challenge of RWD are biases and confounding. In RCTs every participant will be recorded at pre-specified follow-up times, and the health status will be defined at baseline based on inclusion and exclusion criteria of the trial whereby the same criteria will be used in verum and control arms. The health status will have no impact on the sampling scheme. In RWD, on the other hand, the health status of a patient is only recorded, if the patient decides to see a physician. Hence, the density of data during follow-up will heavily depend on the health status. Medication prescriptions in RWD will reflect the treatment guidelines to a large extent. This means that, whenever treatment effects are studied it will be challenging to find suitable control patients in RWD as those patients likely receive non-guideline compliant care. In other words, in any analysis that seeks to establish a causal relationship between treatment and outcome a naively conducted analysis based on RWD will yield heavily confounded results in the vast majority of cases. Also drop-out rates in RWD can be highly biased. For example, in so-called claims data (RWD data which is collected by health insurance companies) a patient drops out of the database if he or she leaves the health insurance plan. In the US, for example, health insurance is often linked to the employment contract, such that unemployment due to health reasons can lead to drop-out from the database. Below we give an example for an apparent treatment (side)-effect that can purely be attributed to confounding.

Example (Obstructive Sleep Apnea After Metformin Prescription)

We are going to investigate potential side-effects of metformin treatment in a cohort of pre-diabetic patients. Metformin is an oral anti-diabetic medication that reduces blood glucose levels in patient with insulin resistance. It is therefore used for the treatment of type II diebetes mellitus (T2DM) but also to prevent progression of prediabetic patients to T2DM. Here we use metformin treatment in a prediabetes cohort for our prototypical analysis. The virtual, i.e. RWD based, trial is constructed using the first pre-diabetes diagnosis as the start of the grace period for treatment initiation. All patients who have already had a diabetes mellitus diagnosis or a metformin prescription before the start of this period are excluded. We now assign patients to either the verum or the control arm in an intention to treat like fashion, i.e. if they initiate metformin within the grace period of three months they are assigned to the verum arm and vice versa. The index time for follow-up then starts after the grace period; i.e. three months after the first pre-diabetes diagnosis.

If we then screen for prevalence of comorbidities in this naively constructed target trial setup we find a significantly higher incidence of obstructive sleep apnea (OSA) in the verum arm compared to the control arm, see Table 3 last row. OSA is condition that causes interruption of normal breathing during sleep. Note that the difference in OSA incidence between verum and control arm represents an association which does not necessarily reflect a causal relationship. Because we have not taken any measures to avoid confounding, treatment and control arm will not be balanced with respect to risk factors for OSA nor will they be balanced with respect to factors that drive Metformin prescription. Hence, it is likely that we can find a confounder, i.e. a parameter that is associated with the treatment decision as well as with the outcome, that accounts for the association between Metformin prescription and OSA

Table 3 Incidence (number of diagnoses per patient year) of obstructive sleep apnea (OSA) after Metformin initiation of pre-diabetic patients

	Number of OSA diagnoses		Patient years [y]		OSA diagnoses per patient year [95% CI]	
BMI	Metformin	No Metformin	Metformin	No Metformin	Metformin	No Metformin
<23.26	11	869	446.4	45,199.4	0.025 [0.011, 0.04]	0.019 [0.018, 0.021]
<25.7	15	1106	677.4	46,444.0	0.022 [0.012, 0.034]	0.024 [0.022, 0.025]
<27.5	26	1527	809.5	46,480.0	0.032 [0.021, 0.044]	0.033 [0.031, 0.035]
<29.16	42	1908	994.7	46,246.9	0.042 [0.03, 0.055]	0.041 [0.039, 0.043]
<30.84	65	2384	1223.2	46,172.5	0.053 [0.041, 0.066]	0.052 [0.05, 0.054]
<32.675	89	2910	1526.2	46,136.8	0.058 [0.047, 0.071]	0.063 [0.061, 0.065]
<34.85	131	3489	1749.5	45,488.2	0.075 [0.062, 0.088]	0.077 [0.074, 0.079]
<37.68	193	4121	2142.1	44,830.5	0.09 [0.077, 0.103]	0.092 [0.089, 0.095]
<42.195	280	5150	2594.2	44,205.2	0.108 [0.096, 0.121]	0.117 [0.113, 0.12]
<97.89	611	7008	3681.3	42,981.7	0.166 [0.153, 0.179]	0.163 [0.159, 0.167]
Total	1463	30,472	15844.4	454,185.2	0.092 [0.088, 0.097]	0.067 [0.066, 0.068]

prevalence that we see in the data. It is well known that obesity is a risk factor for OSA and it hence appears obvious that the body mass index (BMI) might be a suitable covariate for resolving this kind of confounding. In Table 3 it can be seen that simple stratification of the cohort into BMI deciles does indeed resolve the confounding and reveals that probably the association that was initially found is not causal. OSA cannot be considered as a side-effect of Metformin treatment but rather obesity is a risk factor for OSA, and obese patients also have a higher risk for developing diabetes and hence have a higher Metformin prescription rate. This can also be clearly seen by the number of patient years covered by the BMI deciles. In the verum arm the number of patient years increases substantially when going from the lowest to the highest BMI decile, consistent with the assumption that patients with higher BMI will have higher Metformin prescription rates. As the above example shows, any method which does not take into account these biases which are ubiquitous in RWD is likely to result in wrong and misleading conclusions. As we will discuss in the next section, this can be specifically a challenge for AI methods.

Last but not least, it should also be noted that RWD typically has the challenge that one cannot generate new data on demand. One only gets to see the data as patients go to the doctor, without any influence on the type of patients one gets to see. This is in contrast to other data domains where AI methods have been proven particularly successful over the last years, for example, for self-driving cars, computer programs that play board games such as AlphaGo, or chatbots such as chatGPT.

4 Methodological Needs and AI

After having discussed the typical questions that we want to address with RWD (see Sect. 2), as well as RWD's challenging and beneficial features (see Sect. 3), we will now discuss the methods, in particular special methodological needs in the context of pharmaceutical R&D with RWD and AI. We will follow the same structure as in Sect. 2, although we notice that this categorization is not unique as often methodological needs are relevant for several types of questions simultaneously.

4.1 Interpretability in Pharmaceutical Research

Recall that questions arising in the context of basic research are often concerned with gaining a general understanding about the medical need, the SoC, risk factors, as well as a general characterisation of a patient population. This implies that answering these questions typically focuses on interpretability of results, as opposed to pure prediction or (parameter) estimation problems. Thereby, we notice that visualization can be an important aspect about interpretation, with the human visual cortex being a highly efficient pattern recognition machine. While AI and ML methods have the advantage of being able to model highly complex and non-linear relationships in the data, they are typically much harder to interpret (and visualize) than simpler methods.

Interpretability is often addressed via some feature importance metric. For ML/AI methods there are many different ways of defining feature importance, each leading to potentially different rankings among the features. Examples are mean decrease in impurity for tree-ensemble methods [5], gradient-based importance for neural network methods [20], as well as approaches which can be applied more generally to a 'black box' algorithm, such as feature permutation methods [7] and Shapely values [27]. The interpretation of these feature importance metrics itselves is often not straight-forward. In contrast, we note that for classical structural parametric models feature importance is usually straightforward to derive and to interpret via the model coefficients. Moreover, while feature importance for AI and ML methods provide some level of interpretability, it does not capture interpretation of the major advantage of AI/ML methods over classical approaches, namely, interpretation of non-linear, complex interactions in the data. We note that there are some examples from ML tree-ensemble methods, which provide such interpretation of interactions, see e.g., [3, 6].

In general, the field of 'Explainable AI' tries to develop approaches to make AI and ML methods more interpretable. However, in the context of most basic research questions addressed with RWD we note that the primary interest in interpretation is with respect to the data generation process and not with respect to some fitted model. Although, at first glance, both seem to be highly connected, especially for RWD there can be a major difference between the two, as we will discuss in the next paragraph on causality. Regarding the interaction approaches for tree-ensemble methods, some recent results which directly target the data-generation process can be found in [4].

In summary, we find that ML and AI methods often still struggle with achieving a high level of interpretability, as required for many RWD questions arising in basic research. However, new approaches are being developed constantly, but they still need to prove themselves in practice.

4.2 Causality and Statistical Uncertainty in Pharmaceutical Development

Recall that classical pharmaceutical development is mainly concerned with analyzing RCTs. Applications of RWD in this context are, for example, to support study planning, but also more directly via virtual control arms and with pure RWD target trials.

The major reason why RCTs are still widely accepted as the gold-standard of pharmaceutical development is because its randomization allows for a mathematically precise derivation of causal relationships (as opposed to only associations) and hence, can provide clear evidence of a causal drug treatment effect. As discussed in the previous section, RWD data are sparse w.r.t. parameters and longitudinal coverage and analyses based on these data are very susceptible to various kinds of biases and confounding. Therefore, in order to apply AI and ML methods in this context, one needs to modify these prediction-focused methods such that they explicitly take confounding and biases into account. Clearly, the complex black-box nature of AI and ML methods make such modifications more difficult than for simpler models. It is therefore not surprising that many causal inference approaches focus on linear models [10]. Nevertheless, although rare, there exist causal approaches for ML methods, e.g., causal random forest [24], which can be very helpful in situations where linear models do not describe the underlying data generation process well.

It should also be noted that some sub-problems in causal analysis do correspond to pure prediction problems. For example, many causal inference approaches involve estimation of the propensity score.[9] As AI and ML methods are typically very strong for pure prediction problems, they can be much more attractive for establishing such a propensity score model than, for example, linear approaches. Also, when dealing with missing data (which is omnipresent in RWD, as discussed in the previous section), AI and ML methods often have the advantage that they can deal with this more directly. For example, tree-ensemble methods can directly incorporate missingness via the splitting rules and hence, are computationally much more efficient than estimating one model per missingness pattern. More generally, AI and ML algorithms can learn complex manifolds for the data embedding that corresponds to the missingness pattern, which is a major advantage compared to simpler linear models. See e.g., [15, 17] for recent results and discussion on this.

Another major aspect in the context of deriving insights from RCTs, especially when it comes to the final presentation of results for health authorities, is that this will require a precise statistical uncertainty quantification, e.g., via p-values and confidence statements. Such statistical guarantees are not readily available for most AI and ML methods. For this reason, traditional analysis plans as requested by health authorities in the context of RCTs are still mainly building on classical statistical methodology. We note, however, that this is also a moving target, with health authorities continuing to adapt to recent methodological developments. Statistical guarantees for ML and AI is a very active field of research, with many new recent approaches, e.g., in the context of conformal inference, see e.g., [25], including statistical guarantees for feature importance (recall the last sub-section), see [8], but also approaches more targeted to

[9] The propensity score often refers to the probability that a doctor prescribes a certain medication based on a patient's covariates.

specific ML methods, such as tree-ensemble methods, see [2]. We expect that these developments will make AI and ML more applicable to the RWD pharmaceutical R&D context in future.

In summary, we find that causality and statistical uncertainty quantification are two methodological aspects which are essential for most RWD questions arising in pharmaceutical R&D and that both aspects are much harder to implement for AI and ML methods than for classical statistical approaches. However, as for the interpretability aspects in the last sub-section, this is a very active field of research and we expect this to become more applicable in future.

4.3 Medical Expertise in Clinical Practice as a Basis for RWD Methodology

In most cases, the specific methodological needs which were discussed in the previous subsections (interpretability, causality, and statistical uncertainty quantification) also apply to typical questions that are addressed with RWD in clinical practice. Recall that this includes, for example, questions in pharmacovigilance and software as a product in clinical practice.

One further crucial aspect, which we did not discuss so far, is that the analysis of RWD typically requires a very high level of medical expertise. When working with AI/ML this is comparatively more important for the pharmaceutical RWD domain than for other data domains. To see this, note that, in general, AI and ML methods can learn complex relationships in the data without specific prior knowledge as input. This is in contrast to classical statistical methods. For example, for a structural model one typically needs domain expertise in order to decide which features and interaction components need to be included in the model. AI and ML methods typically do not need such input, as they can learn non-linear relationships from the data alone. However, this is often not true in the RWD pharmaceutical context, where one has to take biases and confounding into account. This highly depends on medical prior knowledge and cannot be learned from the data alone. For example, prior knowledge is needed to judge whether a covariate is a confounder in a causal inference setting—something which is essential for bias correction in RWD but which is not a property of the data distribution itself; recall the Metformin example from Sect. 2, where the covariate 'BMI' (obesity) was a confounder.

In summary, we find that there are various methodological needs which occur in the context of pharmaceutical R&D with RWD, which are much more challenging for AI and ML methods than for classical statistical approaches. On the other hand, many of these deficiencies are addressed in an active field of research with promising new approaches being developed in recent times.

5 Summary

Understanding the patient is key to pharmaceutical innovation. Rigorous and quantitative understanding of disease courses, their heterogeneity, influencing factors, and standard of care in real world health care settings is a pillar of pharmaceutical R&D. While traditionally the RCT is the gold standard of clinical exploration and pivotal, confirmatory evidence generation, RWD especially EHR and claims data gains relevance in complementing, enhancing, and in some cases replacing clinical study data. Its potential is explored by the pharmaceutical industry and acknowledged and supported by health authorities worldwide. Applications in pharmaceutical industry range from research, where the exploration of disease specifics, their modifiable properties and the associated true medical need is in focus, over development, where RWD is used to inform the design of clinical trials or provides virtual elements like virtual control arms, up to the post marketing phase where the real world use of pharmaceuticals is explored.

RWD is obviously predestined for AI analyses due to its potential 'big data' nature in terms of length, i.e. number of patients included, as well as breadth, i.e. number of features (diagnoses, prescriptions, interventions, biomarkers etc) covered. Still, RWD comes with some peculiar features that need special methodological consideration and caution. RWD is profoundly sparse and biased. Patients do not see their doctors in a regular pattern and doctors obviously do not perform all available diagnostics regardless of health state at each visit. In turn, healthy patients or patient conditions are under-represented and medical interventions are only applied to (predominantly severely) diseased patients. Additionally, even large data sets only provide little data when it comes to rare medical diseases or conditions. Even for common diseases, dense longitudinal data on e.g. specific markers are rare.

While AI excels in dealing with sparse data, it is highly susceptible to data biases when applied naively. Conventional statistical modeling frameworks provide theoretically well founded means to account for the latter. Substantial medical domain knowledge is needed to inform causality preserving prior structures when designing these models. An extreme example in this regard is a hybrid consisting of a systems pharmacology model which incorporates comprehensive pharmacological and medical knowledge in combination with AI-enabled components [19].

AI is clearly a key enabler for prediction engines in the field of medical devices, they can be trained and operated in a black box fashion if careful validation is provided. In contrast, applications of RWD analyses in pharmaceutical R&D often are about gaining scientific insights rather

than providing a means for predictions. Interpretability of analysis results including their (un-)certainty is a prerequisite in these cases. Conventional statistical models provide confidence (or credibility) ranges and are often easy to interpret by design, as model parameters have a specific meaning in many cases. Still they are limited in complexity by their predefined structure, which might be limited by prior knowledge or prejudice. AI in turn is capable of accurately capturing highly complex, even unexpected, relationships but understanding and interpreting of respective models beyond simple simulations (explainable AI) is still an emerging field, as the assessment of uncertainty is.

Overall, reliable, accurate, and interpretable RWD analyses are indispensable for the ambition to discover and develop differentiated precision medicines. While conventional statistical modeling approaches are fit-for-purpose or even the method of choice in some if not many applications, AI provides a powerful tool to uncover complex unknowns. Its current conceptual limitations will be a matter of future method innovation in the field.

Author Contributions All authors wrote and reviewed the manuscript. CD conducted the analysis of the data examples.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability Statement Certain data used in this study were supplied by International Business Machines Corporation (IBM) as part of one or more IBM Explorys Therapeutic Datasets Delivered. Any analysis, interpretation, or conclusion based on these data is solely that of the authors and not International Business Machines Corporation. The data shown has exemplary character only and cannot be made available for contractual reasons.

Declarations

Conflict of interest CD, JL and RB are employees of Bayer AG and make use of RWD analyses as part of their professional roles.

References

1. Anderson M, Naci H, Morrison D, Osipenko L, Mossialos E (2019) A review of NICE appraisals of pharmaceuticals 2000–2016 found variation in establishing comparative clinical effectiveness. J Clin Epidemiol 105:50–59
2. Athey S, Julie T, Stefan W (2019) Generalized random forests. Ann Stat 47(2)
3. Basu S, Kumbier K, Brown JB, Bin Yu (2018) Iterative random forests to discover predictive and stable high-order interactions. Proc Natl Acad Sci 115(8):1943–1948
4. Merle Behr Yu, Wang XL, Bin Yu (2022) Provable Boolean interaction recovery from tree ensemble obtained via random forests. Proc Natl Acad Sci 119(22):e2118636119
5. Breiman L (2001) Random forests. Mach Learn 45:5–32
6. Bénard C, Biau G, Da Veiga S, Scornet E (2021) SIRUS: stable and interpretable RUle set for classification. Electron J Stat 15(1)
7. Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res 20:1–81
8. Gan L, Zheng L, Allen GI (2022) Inference for interpretable machine learning: fast, model-agnostic confidence intervals for feature importance. arXiv:2206.02088 [cs, stat]
9. Gochanour E, Jayasekera C, Kowdley K (2020) Primary sclerosing cholangitis: epidemiology, genetics, diagnosis, and current management. Clin Liver Dis 15(3):125–128
10. Hernan MA, Robins JM (2023) Causal inference: what if. Chapman & Hall/CRC, Boca Raton
11. Hernán MA, Robins JM (2016) Using big data to emulate a target trial when a randomized trial is not available: table 1. Am J Epidemiol 183(8):758–764
12. Hirschfield GM, Karlsen TH, Lindor KD, Adams DH (2013) Primary sclerosing cholangitis. The Lancet 382(9904):1587–1599
13. Liu F, Demosthenes P (2022) Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 22(1):287
14. Makady A, de Boer A, Hillege H, Klungel O, Goettsch W (2017) What is real-world data? A review of definitions based on literature and stakeholder interviews. Value Health 20(7):858–865
15. Mayer I, Sverdrup E, Gauss T, Moyer J-D, Wager S, Josse J (2020) Doubly robust treatment effect estimation with missing attributes. arXiv:1910.10624 [stat]
16. Merkelbach K, Schaper S, Diedrich C, Fritsch SJ, Schuppert A (2023) Novel architecture for gated recurrent unit autoencoder trained on time series from electronic health records enables detection of ICU patient subgroups. Sci Rep 13(1):4053
17. Morvan M Le, Josse J, Scornet E, Varoquaux G (2021) What's a good imputation to predict with missing values? arXiv:2106.00311 [cs, stat]
18. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL (2010) How to improve R &D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discovery 9(3):203–214
19. Ramaswamy R, Wee SN, George K, Ghosh A, Sarkar J, Burghaus R, Lippert J (2021) CKD subpopulations defined by risk-factors: a longitudinal analysis of electronic health records. CPT: Pharmacom Syst Pharmacol 10(11):1343–1356
20. Shrikumar A, Greenside P, Kundaje A (2017) Learning important features through propagating activation differences. Proc Mach Learn Res 70:3145–3153
21. Stevens PE (2013) Evaluation and management of chronic kidney disease: synopsis of the kidney disease: improving global outcomes 2012 clinical practice guideline. Ann Int Med 158(11):825
22. Strayhorn JM (2021) Virtual controls as an alternative to randomized controlled trials for assessing efficacy of interventions. BMC Med Res Methodol 21(1):3
23. Thorlund K, Dron L, Park JJH, Mills EJ (2020) Synthetic and external controls in clinical trials - a primer for researchers. Clin Epidemiol 12:457–467

24. Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc 113(523):1228–1242
25. Wasserman L, Ramdas A, Balakrishnan S (2020) Universal inference using the split likelihood ratio test. arXiv:1912.11436
26. Zong N, Wen A, Moon S, Fu S, Wang L, Zhao Y, Yu Y, Huang M, Wang Y, Zheng G, Mielke MM, Cerhan JR, Liu H (2022) Computational drug repurposing based on electronic health records: a scoping review. NPJ Digital Med 5(1):77
27. Štrumbelj E, Kononenko I (2014) Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst 41(3):647–665