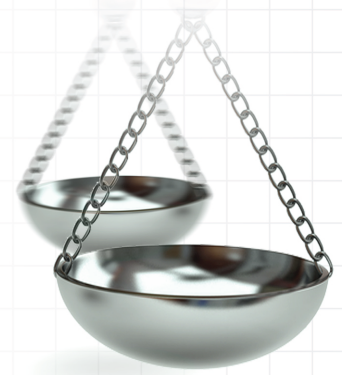




Evaluating the robustness of an AI pathfinder application on eligibility criteria in multiple myeloma trials using real-world data and historical trials

Journal of **Comparative Effectiveness Research**

Rana Jreich^{*1} , Hao Zhang³, Zhaoling Meng² & Fei Wang³

¹R&D, Evidence Generation & Decision Sciences, Sanofi, France

²R&D, Evidence Generation & Decision Sciences, Sanofi, USA

³Department of Population Health Sciences, Weill Cornell Medical College, Cornell University, USA

*Author for correspondence: rana.jreich@sanofi.com

Background: Eligibility criteria are pivotal in achieving clinical trial success, enabling targeted patient enrollment while ensuring the trial safety. However, overly restrictive criteria hinder enrollment and study result generalizability. Broadening eligibility criteria enhances the trial inclusivity, diversity and enrollment pace. Liu *et al.* proposed an AI pathfinder method leveraging real-world data to broaden criteria without compromising efficacy and safety outcomes, demonstrating promise in non-small cell lung cancer trials. **Aim:** To assess the robustness of the methodology, considering diverse qualities of real-world data and to promote its application. **Materials/Methods:** We revised the AI pathfinder method, applied it to relapsed and refractory multiple myeloma trials and compared it using two real-world data sources. We modified the assessment and considered a bootstrap confidence interval of the AI pathfinder to enhance the decision robustness. **Results & conclusion:** Our findings confirmed the AI pathfinder's potential in identifying certain eligibility criteria, in other words, prior complications and laboratory tests for relaxation or removal. However, a robust quantitative assessment, accounting for trial variability and real-world data quality, is crucial for confident decision-making and prioritizing safety alongside efficacy.

First draft submitted: 31 October 2023; Accepted for publication: 17 May 2024; Published online: 13 June 2024

Keywords: AI pathfinder • eligibility criteria • multiple myeloma • RCT replication • real-world evidence

Enhancing clinical trial patient diversity, trial result generalizability and enrollment pace presents a challenge for clinical trial researchers and pharmaceutical industry but is nevertheless encouraged by healthcare regulatory agencies. Eligibility criteria (EC) play a key role in clinical trial design and help researchers achieve targeted and meaningful results. However, recent studies have pointed out that overly restrictive, and sometimes poorly justified ECs slowed down patient enrollment in clinical trials [1]. The US National Cancer Institute concluded that ECs can arbitrarily eliminate patients who could potentially benefit from treatment and thus advocated for their simplification and broadening [2]. For oncology studies, many recommendations and guidance on this topic have been provided by the Food and Drug Administration (FDA), the American Society of Clinical Oncology (ASCO) and Friends of Cancer Research [3,4]. Some recommendations encourage the modernization of ECs related to washout periods, concomitant medications, prior therapies, laboratory reference ranges and test intervals. In April 2022, FDA issued a draft guidance on Diversity Plans to improve enrollment, emphasizing early collaboration between sponsors and agencies to foster more inclusive trial participation [5]. Proactively seeking a better understanding of EC's impact can optimize patient enrollment through more diverse inclusion, while ensuring not to negatively impact trial outcomes, particularly safety, thereby upholding ethical standards.

Real-world data (RWD) is uniquely positioned to provide quantitative assessment of EC's removal or relaxation impact on the trial outcomes. Indeed, the RWD population is heterogeneous and more inclusive since data is collected from everyday clinical practice, outside of the clinical trial setting. Utilizing RWD to broaden clinical trial ECs enables the trial outcomes to more accurately mirror the safety and efficacy of the treatment within

real-world populations. Broadening clinical trial eligibility criteria without negatively impacting study outcomes may enhance the applicability of real-world findings in guiding the selection of eligibility criteria for future clinical trials.

In this context, Liu *et al.* introduced an artificial intelligence algorithm (AI pathfinder) that was applied to real-world data from Flatiron Health and aimed at systematically quantifying the impact of relaxing or removing trial eligibility criteria on non-small cell lung cancer trials and their survival outcomes [6]. Additionally, Rogers *et al.* utilized electronic health records (EHRs) to evaluate the impact of ECs on patient enrollment and safety through the analysis of various combination criteria [7]. Real-world evidence (RWE) derived from RWD offered valuable insights into the most restrictive eligibility criteria, shed light on their high exclusion rates and the potential for broader inclusion to design more inclusive trials and expedite enrollment.

In this paper, we assessed the applicability and robustness of the AI pathfinder introduced by Liu *et al.* in targeting relapsed and refractory multiple myeloma (RRMM). Additionally, we developed and implemented a bootstrapped version of the algorithm to address uncertainties surrounding estimates. Concurrently, we deployed the algorithm across two distinct real-world data sources were deployed to enhance our understanding of real-world outcomes and to evaluate disparities, limitations and data quality for fit-for-use for the disease of interest.

Methods

Data source

Nine historical phase III studies targeting the RRMM patient population with similar disease severity and randomized to two comparable treatments with regards to ECs were identified through ClinicalTrials.gov search and selection by the clinical trial team (Figure 1). The list includes studies CANDOR [8], POLLUX [9], CASTOR [10], APOLLO [11], ASPIRE [12], TOURMALINE [13], ENDEAVOR [14], ELOQUENT-2 [15], – and NIMBUS [16]. We used two real-world data sources (Flatiron Health and Optum's EHR) to investigate the impacts of ECs on these 9 clinical trials.

Flatiron Health database is a nationwide longitudinal EHR-derived database, comprising de-identified patient-level structured and unstructured data, curated via technology-enabled abstraction [17,18]. Most of the patients in the database originate from community oncology settings; relative community/academic portions may vary depending on the study cohort. The de-identified data originated from approximately 280 US cancer clinics (~800 sites of care). The data are de-identified subject to obligations to prevent re-identification and to protect patient confidentiality.

Optum Humedica (EHR) is a data acquisition model that aggregates de-identified EHR data from providers across the continuum of care. Optum's EHR repository is derived from dozens of healthcare provider organizations in the US, that includes more than 57 contributing sources and 111,000 sites of care [19].

RRMM cohort selection

Real-world multiple myeloma (MM) patients were identified using the criteria that required a relevant MM diagnosis ICD-9 203.0x or ICD-10 C90.0x, and these codes appeared in at least two clinical encounters on or after 1 January 2011. The cut-off dates for Flatiron Health and Optum's EHR were February 2022 and June 2020, respectively. Patients with a gap of more than 90 days between the first MM diagnosis and the first encounter in the database were excluded to ensure the diagnosis was confirmed. Relapsed and refractory status cannot be directly derived from RWD due to the lack of information availability. Thus, a proxy of patients with at least 2 lines of MM medications (2L+) was selected as the target of RRMM population. An algorithm was developed to ascertain lines of therapy (LoT) within the Optum's EHR, utilizing prescription records and reported administrations. For each 28-day interval following the initiation of a new treatment regimen, the LoT was updated if a new treatment was documented, or discontinuation occurred. Line 0 was designated for transplant or maintenance therapy lasting at least 60 days, as the initial treatment was not classified as a LoT. Additionally, monotherapy with steroids as the first line was categorized as Line 0. The start of the first LoT was considered to occur after or up to 14 days before the initial MM diagnosis. In the Flatiron Health database, LoT information was available and derived based on a predefined set of disease-specific rules. Real-world patient cohorts were further selected to emulate each of the 9 RRMM RCTs. Target trial emulation required the replication of RCT design principles including ECs, treatment groups, outcomes of interest, etc. [20]. The cohort index date was defined as the starting date of the 2L+ treatment of interest, mimicking the first drug intake in a trial. Clinical records preceding the index date were leveraged to evaluate the computable phenotyping of the original RCT eligibility within RWD. Patient records within a 60-day

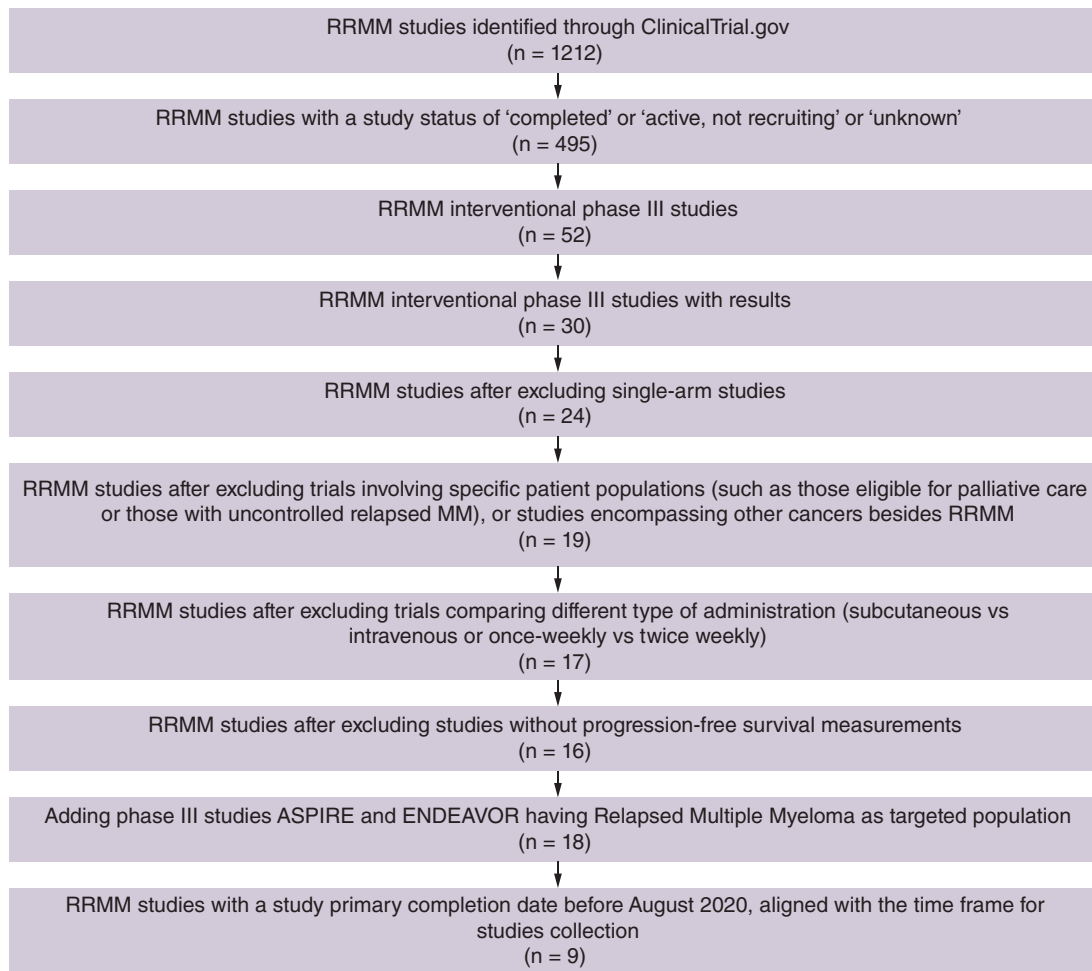


Figure 1. Prisma diagram illustrating the process of identifying the nine relapsed/refractory multiple myeloma studies through ClinicalTrials.gov.
RRMM: Refractory multiple myeloma.

window before the index were employed to emulate baseline characteristics of the RW cohorts. Further elaboration on the derivation of real-world eligibility assessments is shown in [Supplementary Table 1](#).

Variables & assessments

Treatment efficacy was assessed using the primary end point in RRMM trials, progression-free survival (PFS). Unlike solid tumors, the clinical standard for assessing progression in MM relies primarily on changes in M protein spike levels in the blood or urine as measured by protein electrophoresis, which is informed by International Multiple Myeloma Working Group (IMWG) criteria and incorporates both abstracted M protein spike lab values as well as structured free light chain (FLC) lab values for patients with MM that are not detected by M protein spike. In Flatiron Health, progression events were derived according to the IMWG criteria [21]. However, Optum's EHR lacks information about M protein and FLC to derive progression events. Therefore, the initiation of the next MM treatment was utilized as a proxy for a progression event. In both databases, the 'post-baseline' follow-up period was defined as the duration from the index date to either progression disease (PD)/death or the end date of the current LoT whichever occurs earlier. [Table 1](#) summarizes the progression event, follow-up period and censoring definitions in Optum's EHR and Flatiron Health databases, respectively.

Statistical analyses

The methodology employed in this paper is rooted in the trial pathfinder algorithm as conceptualized- by Liu *et al.* [6]. The algorithm workflow began with the emulation of each of the 9 historical RRMM clinical trials,

Table 1. Real-world progression-free survival outcome derivation in Optum’s EHR and Flatiron Health.		
PFS derivation	Optum’s EHR	Flatiron Health
Progression date	Not available. Date of the subsequent line	Date of progression as derived from abstracted M spike or FLC
Definition	Time from index date to next treatment or death	Time from index date to either date of progression or death
Censoring	Last active date	Earliest date between last laboratory test date and end date of current LoT
EHR: Electronic health record; FLC: Free light chain; LoT: Line of therapy; PFS: Progression-free survival.		

which included computable phenotyping implementation of eligibility criteria, generation of treatment groups and calculation of the efficacy outcome of interest, especially the PFS hazard ratio (HR) comparing the treatment of interest against the control group.

The pathfinder quantified the change in PFS treatment HR resulting from the exclusion of specific eligibility criteria across various combinations by employing Shapley values. The method was independently applied to each of the nine historical RRMM trials and considered each eligibility criterion separately. Assuming that N is the set of all computable ECs for a study, the Shapley value of the i-th criterion can be written as:

$$Shapley\ value_i = \sum_{S \in N \setminus \{i\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} (PFS\ HR(S \cup i) - PFS\ HR(S))$$

Where, S is the set of EC combinations without the criterion i, |S| and |N| represent the size of the set S and N, respectively. HR(S) represents the HRs obtained based on the real-world cohort selected by eligibility criteria presented in the set S. N\{i} denotes the set N of all computable ECs without the i-th criterion. A Shapley value close to 0 suggests that criterion i has no significant effect on PFS HR. Conversely, a negative Shapley value indicates that removing the criterion would decrease PFS HR, thereby worsening efficacy.

To address potential selection bias and confounding within real-world cohort treatment groups, we estimated the real-world PFS HR using a weighted Cox proportional hazard model. These weights were derived using the stabilized inverse probability of treatment weighting (sIPTW) method [22]. To address uncertainties surrounding parameter estimates, we implemented a bootstrapped version of the AI pathfinder. Instead of applying the algorithm to the entire fully relaxed RRMM cohort for each trial, we generated 1,000 bootstrapped datasets for each trial, employing replacement. Subsequently, the algorithm was independently applied to each dataset, providing 1000 estimations of Shapley values instead of just one. A confidence interval (CI) was constructed for each eligibility criterion by bootstrapping, enabling the assessment of variability around estimates. Higher uncertainties are observed for CIs that encompass a value of 0. The mean of 1000 bootstrapped values were used in the decision vote. Additionally, the decision rule regarding whether to remove an eligibility criterion was not solely based on Shapley values but was also extended to consider the evidence derived from changes in the size of the real-world data, requiring at least a 5% exclusion rate of patients due to the criterion. The threshold for Shapley values was also adjusted to be equal to or higher than -0.01 (compared with the threshold of 0 used in the original algorithm) to prevent the relaxation of a criterion that might degrade the efficacy response. As the algorithm was independently applied to each of the 9 historical RRMM trials and compared across two different real-world databases, majority voting was utilized to recommend the status of ECs globally. The total number of votes varies for each criterion based on its assessments across the 9 historical RRMM RCTs and applicability across the 2 real-world databases.

Results
Real-world cohort selection & patient count excluded by ECs

A fully relaxed real-world cohort, comprising adult patients with RRMM, was derived, resulting in 13,982 patients in the Optum’s EHR database compared with 6915 in the Flatiron Health database (Figure 2, left panel). Further details on the baseline characteristics of the fully relaxed cohort are provided in Supplementary Table 2. In Flatiron Health, the highest patient exclusions were attributed to Eastern Cooperative Oncology Group (ECOG), prior proteasome inhibitor (PI) and Lenalidomide usage and creatinine clearance, while in Optum’s EHR, the primary reasons for patient exclusion are prior PI and lenalidomide usage, prior asthma/COPD and prior cardiac issues (Figure 2, right panel). Missing data presents a prevalent challenge in the analysis of real-world data. For instance, within the fully relaxed Flatiron Health cohort, approximately 40% of patients exhibited missing ECOG values

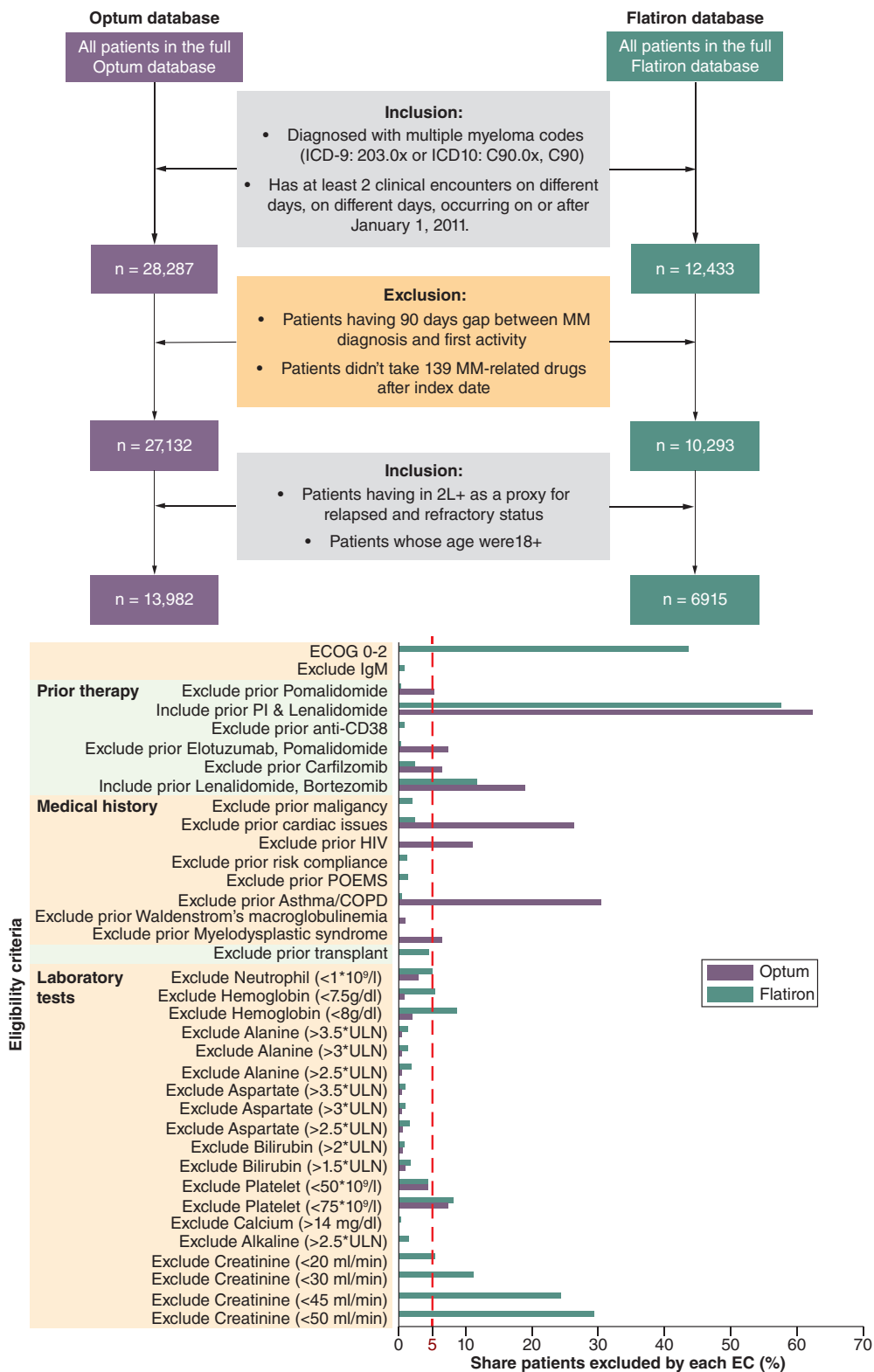


Figure 2. Relapsed and refractory multiple myeloma selection cohorts derived from both Optum's electronic health records and Flatiron Health databases (left panel). Percentage of patients excluded by each of eligibility criterion identified across the ten historical relapsed and refractory multiple myeloma studies (right panel). COPD: Chronic obstructive pulmonary disease; EC: Eligibility criteria; ECOG: Eastern Cooperative Oncology Group; PI: Proteasome inhibitor.

Table 2. The number of eligibility criteria, the number of eligible patients and the hazard ratio of the progression-free survival of the original clinical trials and emulated RRMM trials with eligibility criteria under three scenarios using Flatiron Health database: the original criteria used in the trial, fully relaxed criteria and robust data-driven criteria (i.e., bootstrap AI pathfinder).

Trial name	Clinical trial			Original trial computable EC			Fully relaxed EC		Robust data-driven computable EC		
	EC (n)	Pat. (n)	HR	EC (n)	Pat. (n)	HR	Pat. (n)	HR	EC (n)	Pat. (n)	HR
CASTOR	34	498	0.39 (0.28–0.53)	18	203	1.08 (0.69–1.71)	442	1.12 (0.82–1.52)	16	237	1.06 (0.70–1.62)
ASPIRE	40	792	0.69 (0.57–0.83)	16	325	0.60 (0.26–1.39)	982	1.26 (0.98–1.62)	16	325	0.60 (0.26–1.39)
ENDEAVOR	42	929	0.79 (0.65–0.96)	16	290	1.19 (0.84–1.69)	803	1.39 (1.12–1.72)	15	303	1.17 (0.82–1.65)
ELOQUENT2	42	646	0.70 (0.57–0.85)	18	317	1.59 (0.96–2.66)	713	1.79 (1.25–2.56)	15	378	1.38 (0.86–2.23)
POLLUX	26	569	0.37 (0.27–0.52)	18	315	0.68 (0.46–1.01)	676	0.83 (0.62–1.11)	17	335	0.62 (0.42–0.93)
NIMBUS	35	455	0.48 (0.39–0.60)	15	113	1.22 (0.63–2.37)	456	1.06 (0.79–1.44)	11	282	0.91 (0.62–1.36)
APOLLO	40	304	0.63 (0.47–0.85)	20	141	0.64 (0.39–1.05)	554	0.70 (0.54–0.90)	17	228	0.47 (0.32–0.70)
CANDOR	58	466	0.63 (0.46–0.85)	18	107	0.94 (0.37–2.38)	358	0.98 (0.66–1.45)	15	135	0.68 (0.28–1.66)
TOURMALINE	27	722	0.74 (0.59–0.94)	14	340	1.15 (0.80–1.67)	695	1.57 (1.18–2.08)	13	354	1.11 (0.78–1.59)

The fully relaxed criteria correspond to evaluating the hazard ratio of PFS of all the patients in the Flatiron Health database who took the treatments in the relevant line of therapy.
EC: Eligibility criteria; HR: Hazard ratio; Pat.: Patient; RRMM: Relapsed and refractory multiple myeloma.

during the baseline period assessment. Furthermore, laboratory tests within the Flatiron Health database showed a range of missing values, from approximately 13 to 34%. To prevent a large reduction in cohort sample size resulting from the application of all inclusion/exclusion criteria, patients with missing ECOG or laboratory test values at baseline were considered eligible for analysis.

EC removal or relaxation across different real-world datasets & trials

Figure 3 showed the results of the bootstrapped pathfinder algorithm for the APOLLO, POLLUX and ASPIRE trials. Similar plots were provided for the remaining 6 RCTs in [Supplementary Figure 1](#). A good consistency was observed between the bootstrapped version of the algorithm and the version without bootstrapping (almost all CIs are located at the left or right of the Shapley values threshold equal to -0.01). The means of bootstrapped Shapley values were considered as the final estimates for decision-making to enhance decision robustness. Higher variability in Shapley values *CI*s was observed in Optum's EHR versus Flatiron Health, which can be explained by the presence of high heterogeneity of patient populations in Optum's EHR compared with the Flatiron Health database. Figure 4 summarized the decision-making process for broadening eligibility criteria across the nine historical RRMM trials. This included criteria to keep (Shapley value < -0.01 and % of patients excluded ≥ 5%), criteria to remove or relax (Shapley value ≥ -0.01 and % of patients excluded ≥ 5%), and criteria with insufficient evidence to make a judgment (less than 5% of RWD patients excluded by the criterion application). Results were derived for both Optum's EHR and Flatiron Health databases. For the same RCT, the EC removal decision could be different between the two RWD sources. This may reflect different real world data qualities, data availability and missing values, etc. For example, based on Flatiron Health, a potential negative impact on efficacy was detected when removing some eligibility criteria such as ECOG 0-2 (8/9 voting to be kept) and the requirement of no more than three regimens for MM (4/5 voting to be kept). Thus, these two criteria should not be removed or relaxed based on Flatiron Health database. However, ECOG values and the requirement of more than three regimens for MM were not available in Optum's EHR to be assessed. On the other hand, based on Optum's EHR, exclusion of patients with prior HIV (5/8 to be removed), or prior cardiac issues (5/8 to be removed), or prior asthma/COPD (3/4 to be removed) should be considered for removal or relaxation. Based on Flatiron Health database, the decisions on these 3 criteria were uncertain, which could be due to higher missing values on prior events in this database. Furthermore, prior therapies are very specific to each clinical study and treatment of interest and thus no judgment can be made across all RRMM trials. Regarding laboratory tests, the rule of majority voting could be very restrictive and strict. For a better understanding of laboratory tests' impact, all labs having at least three potential removals or relaxation detections across the 9 historical RRMM trials and the 2 real-world databases should be investigated for removal or more probability for eligible threshold relaxation, such as Neutrophil, Creatinine Clearance and Platelet.

Table 2 & [Supplementary Table 3](#) summarized the PFS HR results obtained on eligible cohorts (RW cohorts

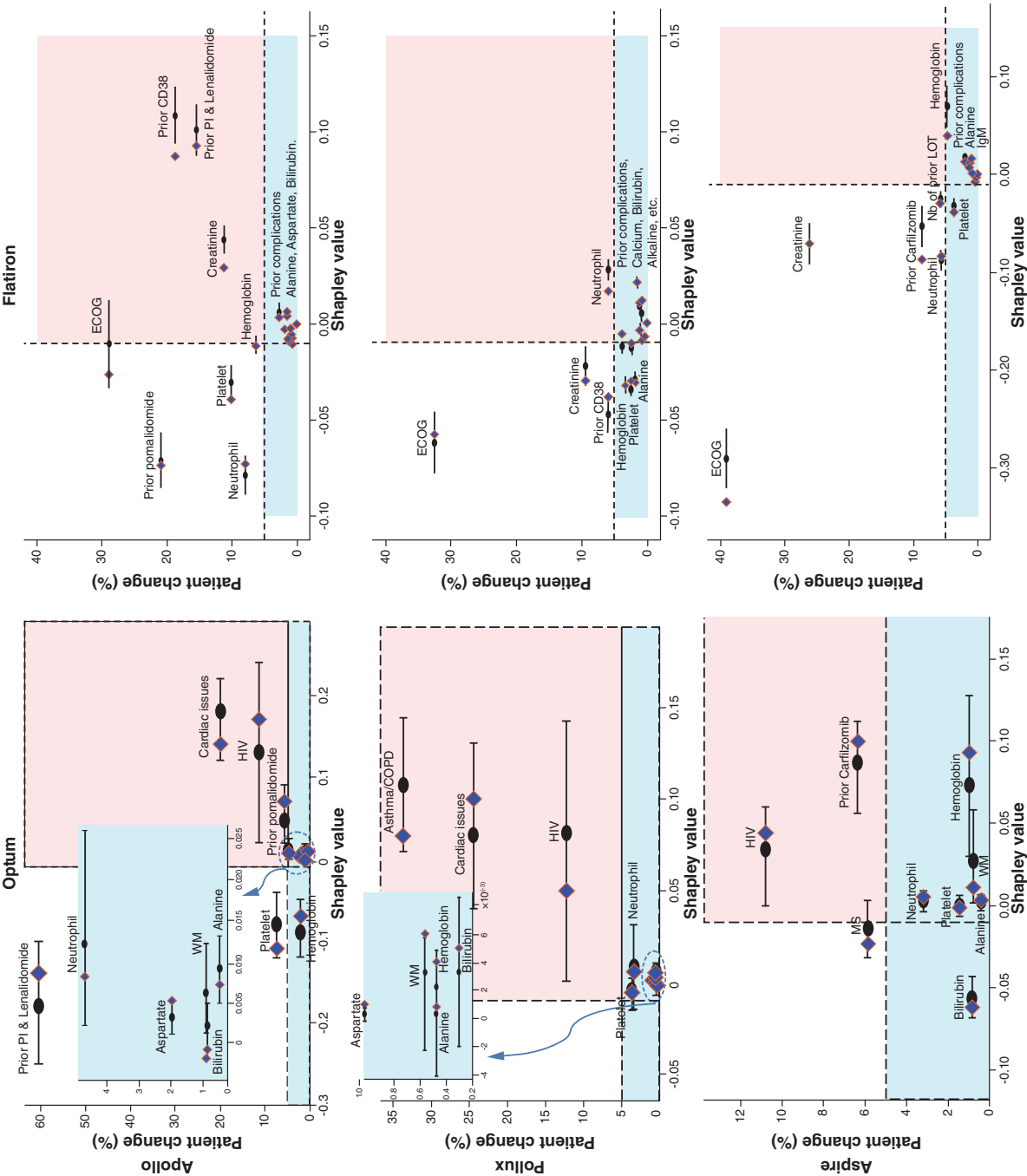


Figure 3. Percentage of excluded patients versus Shapley values by each eligibility criterion for APOLLO, POLLUX and ASPIRE trials using both Optum's electronic health records and Flatiron Health real-world databases. Bootstrapped algorithm results are showed using confidence intervals. The black points represent the mean of bootstrapped Shapley values. The blue rhombus indicates results obtained by the same algorithm without bootstrapping. Eligibility criteria located in the blue zone led to less than 5% of real-world patient count exclusion, while the orange zone indicates eligibility criteria eligible for relaxation or removal (with both Shapley value and percentage of patients excluded higher than -0.01 and 5%, respectively).
ECOG: Eastern Cooperative Oncology Group; LOT: Line of therapy; PI: Proteasome inhibitor; WM: Waldenström's macroglobulinemia.

	Trial name	APO.		POL.		CAS.		ASP.		CAN.		NIM.		END.		ELO2		TOUR.		Flatiron health		Optum's EHR		Decision making				
	Database ECOG 0-2	O	F	O	F	O	F	O	F	O	F	O	F	O	F	O	F	O	F	Keep (8/9) Remove (1/9)	NA	NA	Keep					
	IGM																			NA	NA	NA	NA					
	Prior transplant																			Uncertain (6/5)	NA	Uncertain	Uncertain					
	<= 3 regimens																			Keep (4/5) Remove (1/5)	NA	Keep	Keep					
Prior therapy	Prior Lenalidomide, Bortezomib (I)																			Trial specific								
	Prior Elicuzumab, Pomalidomide																											
	Prior PI & Lenalidomide (I)																											
	Prior Pomalidomide																											
Medical history	Prior anti-CD38																											
	Prior HIV																	*		Uncertain (9/9)	Keep (3/8) Remove (5/8)	Remove	Remove					
	Prior risk compliance																			Uncertain (9/9)	NA	Uncertain	Uncertain					
	Prior asthma/COPD																			Uncertain (4/4)	Remove (3/4) Uncertain (1/4)	Remove	Remove					
	Prior cardiac issues																	*		Uncertain (9/9)	Keep (2/8) Remove (5/8) Uncertain (1/8)	Remove	Remove					
	Prior POEMS or Amyloidosis																	*		Uncertain (8/8)	NA	Uncertain	Uncertain					
	Prior malignancy																			Uncertain (6/6)	NA	Uncertain	Uncertain					
	Prior WM																			NA	Uncertain (1/2) Remove (1/2)	Uncertain	Uncertain					
Lab test	Prior MS																			NA	Uncertain (8/8)	Uncertain	Uncertain					
	Bilirubin																	*		Uncertain (9/9)	Uncertain (8/8)	Uncertain	Uncertain					
	Creatinine																	*		Keep (5/9) Remove (3/9) Uncertain (1/9)	NA	Investigate threshold relaxation	Investigate threshold relaxation					
	Calcium																	*		Uncertain (7/7)	NA	Uncertain	Uncertain					
	Aspartate																	*		Uncertain (7/7)	Uncertain (5/6) Remove (1/6)	Uncertain	Uncertain					
	Alkaline																			Uncertain (1/9)	NA	Uncertain	Uncertain					
	Alanine																	*		Uncertain (9/9)	Uncertain (9/9)	Uncertain	Uncertain					
	Hemoglobin																	*		Keep (5/9) Uncertain (4/9)	Uncertain (8/8)	Keep	Keep					
	Neutrophil																	*		Remove (6/9) Keep (3/9)	Uncertain (8/8)	Remove	Remove					
	Platelet																	*		Keep (4/9) Remove (3/9) Uncertain (2/9)	Keep (4/8) Uncertain (4/8)	Investigate threshold relaxation	Investigate threshold relaxation					
Color illustration	Not applicable																											

Figure 4. Summary of eligibility criteria broadening decision across the 9 historical Relapsed and refractory multiple myeloma trials using Optum's electronic health records and Flatiron Health. Decision making was based on the number of patients change and Shapley values. Real-world databases Optum's EHR and Flatiron Health are referred as O and F respectively. * Indicates trial with presence of trend efficacy result between RCT and RWE studies.
APO.: APOLLA; ASP.: ASPIRE; CAN.: CANDOR; CAS: CASTOR; COPD: Chronic obstructive pulmonary disease; ECOG: Eastern Cooperative Oncology Group; EHR: Electronic health record; ELO2: ELOQUENT; END.: ENDEAVOR; IGM: Immunoglobulin; MS: Myelodysplastic syndrome; NIM.: NIMBUS; PI: Proteasome inhibitor; POEMS: Polyneuropathy, organomegaly, endocrinopathy/edema, monoclonal-protein and skin changes; POL.: POLLUX; TOUR.: TOURMALINE; WM: Waldenström's macroglobulinemia.

that respect both treatment groups and RCTs original I/E criteria), non-eligible cohorts (RW cohorts that respects treatment groups but not RCTs I/E criteria) and data-driven cohorts (RW cohorts that respects treatment groups and RCTs I/E criteria that should not be relaxed based on AI pathfinder) using Optum's EHR and Flatiron Health. In Flatiron Health database, the AI pathfinder resulted in the removal of an average of 2 criteria across the nine historical RRMM RCTs, with an average change of 0.12 in PFS HR and an increase of 47-patients in the real-world cohort. Similarly, in Optum's EHR, the AI pathfinder led to an average removal of 2 criteria across the 9 evaluated RCTs, accompanied by an average change of 0.20 in TTNT HR and a gain of 178 patients on average. Note the identified ECs could be different based on two different databases.

Limitations

Optum's EHR database could be less well-positioned for oncology PFS assessment, especially in the absence of ECOG information and progression events. While Flatiron Health served well for PFS analysis, it lacked comprehensive patient medical history, especially concerning the timing of data capture. Flatiron Health data for many patients were only recorded upon their initial follow-up with an oncologist in Flatiron's Health OncoEMR, leading to inconsistent availability of previous medical records for all individuals. Both Flatiron Health and Optum's EHR databases suffered from important amount of missing laboratory test values at baseline, which refrained the robustness of studying the impact of laboratory tests at baseline on the study outcome. A higher than expected percentage of patient exclusion due to prior cardiac issues, prior Asthma/COPD and prior HIV was observed in Optum's EHR, highlighting the necessity for better computable phenotyping approaches for these conditions to specifically target those indicating significant events. One might question the feasibility of employing the developed AI pathfinder to evaluate the collective impact of a subset of ECs rather than assessing each criterion individually. This alternative approach could offer greater robustness by capturing the interplay among ECs more effectively. Additionally, concerns may arise regarding the applicability of these findings to inform and optimize future RRMM studies, particularly in instances of discordance between efficacy results from RCTs and RWE studies. Notably, the replication of PFS HR analysis using the Flatiron Health database yielded significant findings for only 3 out of the 9 replicated RRMM studies. While using Optum's EHR, only one study exhibited concordant efficacy results between RCT and RWE. The disparities should be noticed in measurement availability and recommendations for EC removal/retention between Flatiron Health and Optum's EHR. Improved real-world data quality is imperative for a more robust and informed assessment of ECs, particularly concerning laboratory tests such as those related to liver function (alanine, aspartate, bilirubin), where insufficient evidence was available from both Optum's EHR and Flatiron Health due to a significant proportion of missing values.

Conclusion

In this paper, we have explored the trial AI pathfinder approach, initially developed by Liu *et al.* [6] incorporating several novel improvements. Firstly, we introduced a bootstrapped version of the algorithm to assess the variability surrounding Shapley value estimates and offer a more robust analysis. Secondly, we evaluated the algorithm's performance across two distinct real-world data sources, aiming to discern the necessary data quality requirements for its effective application. Lastly, this investigation extended the algorithm's application to a new therapeutic area, focusing on RRMM. The aim was to broaden its utility and demonstrate its adaptability to diverse disease domain. Independently applied to nine historical RCTs, the AI pathfinder identified opportunities for the potential removal or relaxation of eligibility criteria related to prior complications, such as HIV, cardiac issues and asthma/COPD, along with the possible relaxation of laboratory test thresholds (Neutrophil, Creatinine Clearance and Platelet). These recommendations were based on Shapley values and the proportion of real-world patients excluded by each criterion, determined through majority voting across the 9 historical RRMM trials and the two distinct real-world data sources. However, different recommendations can be reached based on different real-world databases, which can be related to different measurement availability and different percentages of missing data between the two databases. The robustness of the decision could be questioned. Further exploration is also necessary to determine whether there is a more effective approach to handling missing values in real-world data, as well as to assess the applicability of AI pathfinder findings in cases where non-concordance efficacy was observed between RCT findings and RWE analysis after emulation.

Furthermore, despite the obtained results, a safety analysis is needed to confirm the final subset of ECs to remove or relax. The final decision-making for ECs removal or threshold relaxation should be derived jointly based on both efficacy and safety analyses. Rogers *et al.* applied a k-means clustering to derive which combinations of different

ECs maximized patients counts but minimized hospitalization risk, leveraging electronic health record data. This approach was applied in three disease domains: relapsed/refractory lymphoma/leukemia, hepatitis C virus and chronic kidney disease. In this context, one may envision leveraging pathfinder for safety analysis. This could involve substituting the alteration in PFS HR between cohorts with and without a particular eligibility criterion with the HR of hospitalization risk or occurrence of adverse events (AEs). However, in generally, RCTs are not powered to detect safety outcomes, especially for less frequent events. Furthermore, the assessment and reporting of safety events differ between real-world data and clinical setting. Hence, further research is warranted to evaluate the impact of ECs on study safety outcomes.

This effort to quantitatively measure the impact of ECs on efficacy and safety outcomes assists in refining the inclusion and exclusion criteria of RCTs, thereby widening the scope of the targeted patient population and accelerating study enrollment. Moreover, following the identification of potential ECs for relaxation based on efficacy and safety analyses, it is imperative to evaluate the broader impact of ECs on sociodemographic factors, such as age, gender, race and ethnicity, and how they may contribute to a more inclusive trial environment.

Summary points

Challenge of clinical trial diversity

- Pharmaceutical industries face challenges in enhancing patient diversity, result generalizability and enrollment, yet are encouraged by regulatory agencies.

Role of eligibility criteria

- Eligibility criteria (ECs) are crucial in trial design, but overly restrictive criteria can hinder patient enrollment, prompting recommendations for their simplification and broadening.

FDA draft guidance on diversity plans

- The FDA proposed a draft guidance encouraging diversity plans to improve enrollment from underrepresented populations, emphasizing early collaboration between sponsors and agencies.

Utilization of real-world data & emulating RCT trials

- Real-world data (RWD) provides a unique opportunity for quantitative assessment of ECs removal or relaxation, enabling trial outcomes to reflect real-world treatment efficacy. However, it's crucial to evaluate the efficacy-effectiveness gap and assess the applicability of different qualities of real-world data in replicating RCT findings within the clinical setting.

Application of AI algorithms

- AI algorithms, such as the one introduced by Liu *et al.*, systematically quantify the impact of ECs on trial populations and survival outcomes, aiding in protocol optimization.

Study focus on relapsed & refractory multiple myeloma

- The study assesses the applicability and robustness of AI algorithms in targeting relapsed and refractory multiple myeloma (RRMM), addressing uncertainties and disparities across real-world data sources.

Methodology for cohort selection

- Real-world patient cohorts are selected to emulate nine historical RRMM clinical trials using two different real-world databases Optum's electronic health record and Flatiron Health.

Assessment of treatment efficacy

- Treatment efficacy is evaluated using progression-free survival as the primary end point, and hazard ratio calculations.

Decision-making & recommendations

- Majority voting is utilized to recommend the status of ECs globally, considering both Shapley values and evidence from real-world data to inform relaxation decisions.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: <https://bpl-prod.literatumonline.com/doi/10.57264/cer-2023-0164>

Financial disclosure

This work was supported by Sanofi. Rana Jreich and Zhaoling Meng are employees of Sanofi and hold shares and/or stock options in Sanofi. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Competing interests disclosure

The authors have no competing interests or relevant affiliations with any organization or entity with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Writing disclosure

No writing assistance was utilized in the production of this manuscript.

Data sharing statement

The data supporting the findings of this analysis were obtained from Flatiron Health and Optum's EHR. These de-identified datasets may be made available upon request and are subject to a license agreement.

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc-nd/4.0/>

References

Papers of special note have been highlighted as: • of interest

1. Van Spall HGC, Toren A, Kiss A *et al.* Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *JAMA* 297(11), 1233–1240 (2007).
2. Scoggins JF, Ramsey SD. A national cancer clinical trials system for the 21st century: reinvigorating the NCI Cooperative Group Program. *J. Natl Cancer Inst.* 102(17), 1371–1371 (2010).
3. Kim ES, Uldrick TS, Schenkel C *et al.* Continuing to broaden eligibility criteria to make clinical trials more representative and inclusive: ASCO-friends of cancer research joint research statement. *Clin. Cancer Res.* 27(9), 2394–2399 (2021).
4. Duggal M, Sacks L, Vasisht KP. Eligibility criteria and clinical trials: An FDA perspective. *Contemp. Clin. Trials* 109, 106515 (2021).
- **Offers a valuable insight from FDA perspective and guidance on eligibility criteria for optimizing trial design and ensuring patient inclusion, safety and regulatory compliance.**
5. FDA services US department of Health and Human. Enhancing the diversity of clinical trial populations-eligibility criteria, enrollment, practices, and trial design guidance for industry. Available at: <https://www.federalregister.gov/documents/2020/11/10/2020-24881/enhancing-the-diversity-of-clinical-trial-populations-eligibility-criteria-enrollment-practices-and#> (2020).
- **Offers the FDA's perspective on the imperative to enhance diversity within clinical trial populations, providing comprehensive guidance on eligibility criteria, enrollment practices and trial design to foster inclusivity and improve the generalizability of study outcomes.**
6. Liu R, Rizzo S, Whipple S *et al.* Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature* 592(7855), 629–633 (2021).
- **Provides a deeper understand on how real-world data and artificial intelligence are used to assess the significance of eligibility criteria in oncology trials.**
7. Rogers JR, Pavisic J, Ta CN *et al.* Leveraging electronic health record data for clinical trial planning by assessing eligibility criteria's impact on patient count and safety. *J. Biomed. Inform.* 127, 104032 (2022).
- **Presents a machine learning approach employed to assess the impact of eligibility criteria on patient enrollment rates and safety using electronic health record data, offering valuable insights into optimizing clinical trial design.**
8. Dimopoulos M, Quach H, Mateos MV *et al.* Carfilzomib, dexamethasone, and daratumumab versus carfilzomib and dexamethasone for patients with relapsed or refractory multiple myeloma (CANDOR): results from a randomised, multicentre, open-label, phase 3 study. *Lancet* 396(10245), 186–197 (2020).
9. Dimopoulos MA, Oriol A, Nahi H *et al.* Daratumumab, lenalidomide, and dexamethasone for multiple myeloma. *N. Engl. J. Med.* 375(14), 1319–1331 (2016).
10. Palumbo A, Chanan-Khan A, Weisel *et al.* Daratumumab, bortezomib, and dexamethasone for multiple myeloma. *N. Engl. J. Med.* 375(8), 754–766 (2016).
11. Dimopoulos MA, Terpos E, Boccadoro M *et al.* Daratumumab plus pomalidomide and dexamethasone versus pomalidomide and dexamethasone alone in treated multiple myeloma (APOLLO): an open-label, randomised, phase 3 trial. *Lancet Oncol.* 22(6), 801–812 (2021).
12. Stewart AK, Rajkumar SV, Dimopoulos MA *et al.* Carfilzomib, lenalidomide, and dexamethasone for relapsed multiple myeloma. *N. Engl. J. Med.* 372(2), 142–152 (2015).
13. Moreau P, Masszi T, Grzasko N *et al.* Oral ixazomib, lenalidomide, and dexamethasone for multiple myeloma. *N. Engl. J. Med.* 374(17), 1621–1634 (2016).

14. Dimopoulos MA, Moreau P, Palumbo A *et al.* Carfilzomib and dexamethasone versus bortezomib and dexamethasone for patients with relapsed or refractory multiple myeloma (ENDEAVOR): a randomised, phase 3, open-label, multicentre study. *Lancet Oncol.* 17(1), 27–38 (2016).
15. Lonial S, Dimopoulos M, Palumbo A *et al.* Elotuzumab therapy for relapsed or refractory multiple myeloma. *N. Engl. J. Med.* 373(7), 621–631 (2015).
16. Moreau P, Weisel KC, Song KW *et al.* Relationship of response and survival in patients with relapsed and refractory multiple myeloma treated with pomalidomide plus low-dose dexamethasone in the MM-003 trial randomized phase III trial (NIMBUS). *Leuk. Lymphoma* 57(12), 2839–2847 (2016).
17. Birnbaum B, Nussbaum N, Seidl-Rathkopf K *et al.* Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. Available at: <https://arxiv.org/abs/2001.09765>
18. Ma X, Long L, Moon S *et al.* Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron Health, SEER, and NPCR. Available at: <https://www.medrxiv.org/content/10.1101/2020.03.16.20037143v3>
- **Provides a deeper understanding of Flatiron Health data heterogeneity and its implications for research and clinical practice.**
19. Optum® de-identified Electronic Health Record dataset (2007–2021).
- **Provides a deeper understanding of OPTUM electronic Health record offering a valuable resource for researchers to analyze long-term trends, outcomes and patterns in healthcare delivery and patient care.**
20. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.* 183(8), 758–764 (2016).
- **Presents a comprehensive framework and decision guide aimed at facilitating the selection of optimal strategies for effectively emulating a clinical trial using real-world data.**
21. Rajkumar SV, Dimopoulos MA, Palumbo A *et al.* International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. *Lancet Oncol.* 15(1), e538–e548 (2014).
22. Xu S, Ross C, Raebel MA *et al.* Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value Health* 13(2), 273–277 (2010).
- **Introduces a statistical approach aimed at reducing confounding in the estimation of relative risk when analyzing real-world data.**