

Article

Reducing Sample Size While Improving Equity in Vaccine Clinical Trials: A Machine Learning-Based Recruitment Methodology with Application to Improving Trials of Hepatitis C Virus Vaccines in People Who Inject Drugs

Richard Chiu ^{1,2} , Eric Tataru ^{3,4}, Mary Ellen Mackesy-Amiti ⁵ , Kimberly Page ⁶ , Jonathan Ozik ^{3,4} , Basmattee Boodram ⁵, Harel Dahari ²  and Alexander Gutfraind ^{2,6,*} 

¹ Department of Medicine, University of Illinois College of Medicine at Chicago, Chicago, IL 60612, USA; rchiu8@uic.edu

² The Program for Experimental & Theoretical Modeling, Department of Medicine, Division of Hepatology, Stritch School of Medicine, Loyola University Chicago, Maywood, IL 60660, USA; hdahari@luc.edu

³ Consortium for Advanced Science and Engineering, University of Chicago, Chicago, IL 60637, USA

⁴ Argonne National Laboratory, Lemont, IL 60439, USA

⁵ Division of Community Health Sciences, School of Public Health, University of Illinois at Chicago, Chicago, IL 60612, USA; bboodram@uic.edu (B.B.)

⁶ Department of Internal Medicine, Division of Epidemiology, Biostatistics and Preventive Medicine, University of New Mexico Health Sciences Center, Albuquerque, NM 87131, USA; pagek@salud.unm.edu

* Correspondence: agutfraind@luc.edu; Tel.: +1-708-216-4682



Citation: Chiu, R.; Tataru, E.; Mackesy-Amiti, M.E.; Page, K.; Ozik, J.; Boodram, B.; Dahari, H.; Gutfraind, A. Reducing Sample Size While Improving Equity in Vaccine Clinical Trials: A Machine Learning-Based Recruitment Methodology with Application to Improving Trials of Hepatitis C Virus Vaccines in People Who Inject Drugs. *Healthcare* **2024**, *12*, 644. <https://doi.org/10.3390/healthcare12060644>

Academic Editor: Tin-Chih Toly Chen

Received: 29 December 2023

Revised: 1 March 2024

Accepted: 6 March 2024

Published: 13 March 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Despite the availability of direct-acting antivirals that cure individuals infected with the hepatitis C virus (HCV), developing a vaccine is critically needed in achieving HCV elimination. HCV vaccine trials have been performed in populations with high incidence of new HCV infection such as people who inject drugs (PWID). Developing strategies of optimal recruitment of PWID for HCV vaccine trials could reduce sample size, follow-up costs and disparities in enrollment. We investigate trial recruitment informed by machine learning and evaluate a strategy for HCV vaccine trials termed PREDICTEE—Predictive Recruitment and Enrichment method balancing Demographics and Incidence for Clinical Trial Equity and Efficiency. PREDICTEE utilizes a survival analysis model applied to trial candidates, considering their demographic and injection characteristics to predict the candidate's probability of HCV infection during the trial. The decision to recruit considers both the candidate's predicted incidence and demographic characteristics such as age, sex, and race. We evaluated PREDICTEE using in silico methods, in which we first generated a synthetic candidate pool and their respective HCV infection events using HepCEP, a validated agent-based simulation model of HCV transmission among PWID in metropolitan Chicago. We then compared PREDICTEE to conventional recruitment of high-risk PWID who share drugs or injection equipment in terms of sample size and recruitment equity, with the latter measured by participation-to-prevalence ratio (PPR) across age, sex, and race. Comparing conventional recruitment to PREDICTEE found a reduction in sample size from 802 (95%: 642–1010) to 278 (95%: 264–294) with PREDICTEE, while also reducing screening requirements by 30%. Simultaneously, PPR increased from 0.475 (95%: 0.356–0.568) to 0.754 (95%: 0.685–0.834). Even when targeting a dissimilar maximally balanced population in which achieving recruitment equity would be more difficult, PREDICTEE is able to reduce sample size from 802 (95%: 642–1010) to 304 (95%: 288–322) while improving PPR to 0.807 (95%: 0.792–0.821). PREDICTEE presents a promising strategy for HCV clinical trial recruitment, achieving sample size reduction while improving recruitment equity.

Keywords: randomized clinical trial; vaccine trial recruitment; hepatitis C; equity; people who inject drugs; machine learning

1. Introduction

Despite over thirty years of research, there is still no vaccine for hepatitis C virus (HCV) infection [1,2], a virus affecting over 2 million adults in the U.S. alone [3,4]. While HCV can be cured with direct-acting antivirals (DAAs) [5], treatment with DAAs can cost upwards of \$25,000 [6] and does not prevent reinfection [7]. A vaccine promises to be an inexpensive, feasible, and accessible viral control strategy, especially for high-risk populations such as people who inject drugs (PWID), who are affected by the opioid crisis [8].

Our primary motivation for improving vaccine clinical trial recruitment is to reduce the sample size needed in an HCV vaccine trial to achieve adequate statistical power. An analysis of randomized clinical trials (RCTs) revealed that 19% were terminated due to insufficient accrual of participants [9], and up to 86% do not reach their recruitment targets within the originally envisaged timescale [10,11]. Moreover, studies place the cost of recruitment to a vaccine clinical trial to be in the range of \$9000 to \$16,000 per recruited subject, considering expenses such as tracking and retention costs, reimbursement, and drug costs [12,13]. Reducing sample size therefore not only increases the feasibility of vaccine trials, but also enables significant cost savings.

We secondarily seek to improve representation of diverse populations in HCV RCTs and ensure equitable access to novel vaccines. A study by Wilder et al. found that participation of Black/African American participants in HCV-related clinical trials does not proportionally reflect the burden of HCV among this demographic in North America [14]. The issue of adequate representation is particularly pertinent in the context of HCV because demographic characteristics such as sex [15–17], race [15,18], and age [19] can impact HCV viral clearance and immunology and likely alter vaccine efficacy. The FDA recently called for innovative strategies promoting enrollment of underrepresented populations in RCTs [20] and the NIH instructed trials it funds to ensure adequate inclusion of minorities [21]. However, it is desired to achieve these objectives without magnifying the costs of screening and the overall sample size.

This study aims to address both sample size and equity concerns by introducing a novel approach for recruiting for HCV vaccine clinical trials. The approach integrates machine learning (ML) to identify persons most likely to be exposed to HCV while also implementing a weighting scheme prioritizing underrepresented candidates. We refer to this recruitment strategy as the Predictive Recruitment and Enrichment method balancing Demographics and Incidence for Clinical Trial Equity and Efficiency (PREDICTEE). We surveyed the literature for similar work and identified several studies that used a predictive model for recruitment of individuals at high risk for the outcome of interest [22–25] as well as others that discuss strategies to improve equity and inclusion in RCTs [26]. There have also been other suggestions for improving trials for HCV vaccine candidates (e.g., [27,28]), but none targeting sample size or equity. Therefore, we believe this study is the first to achieve the two opposing objectives in the context of vaccine trial recruitment for HCV vaccines.

This study contributes to a growing body of literature on the implications of artificial intelligence (AI) and machine learning in clinical trials. Recent reviews by Ismail et al. [29] and Harrer et al. [30] describe how AI/ML may be leveraged to optimize recruitment to reduce trial failures, optimize patient composition, and reduce time and costs associated with conducting an RCT. We believe PREDICTEE offers one such ML-based solution to these goals. This study also complements a recent study by Oikonomou et al. [31] which describes a ML-based phenomapping strategy capable of maximizing RCT enrollment efficiency that is similar to what we propose in our PREDICTEE methodology. However, we offer an alternative method of optimizing the clinical trial cohort and apply it to a simulated vaccine clinical trial.

PREDICTEE also builds upon our previous work on a model of Hepatitis C elimination in PWID (HepCEP), which simulates HCV infection, network formation, and syringe sharing in the PWID population of metropolitan Chicago [32]. The present study investigates how longitudinal PWID data, such as that generated using the HepCEP model, can be

leveraged to improve vaccine trial recruitment equity and efficiency among PWID. We also propose a novel use of survival analysis in clinical trial recruitment decisions for the purpose of prognostic enrichment. Using simulations of HCV vaccine trial recruitment of PWID based on data collected from our previous study, we show that PREDICTEE yields trial cohorts that are approximately half the size of those recruited using conventional methods, while also improving demographic representativeness.

2. Materials and Methods

In designing PREDICTEE, we set multiple objectives: (1) to identify subgroups with high incidence in the trial cohort while ensuring that (2) the final sample satisfies constraints on representation. Additionally, the recruitment process should (3) be efficient and conclude in a pre-specified amount of time and (4) cope with uncertainty about the quality and background of trial candidates. Below, we describe how PREDICTEE works and evaluate its performance in a simulated HCV vaccine RCT in Chicago.

2.1. Preparation

2.1.1. Longitudinal Data

Our longitudinal data on PWID are derived from a large synthetic PWID population. To achieve this, we used the HepCEP agent-based model that simulates PWID behavioral patterns—daily injection drug use, social network formation and dissolution, and geography [32,33]. Details of the HepCEP model are described in the Supplementary Materials; in brief, the HepCEP model simulates events such as PWID attrition, new PWID arrival, drug sharing, network formation, HCV infection, recovery, vaccination and more. To generate our synthetic population, we constrained the HepCEP model to maintain a population of 100,000 PWID over its 10-year simulation, removed any simulated vaccine effects, and kept all other HepCEP parameters at default. This generated profiles for 123,071 PWID—the pool of candidates we used for our vaccine trial recruitment. Each PWID profile contains demographic, network, and injection characteristics, as well as the recorded time to HCV infection if it occurred. Additional variables were also calculated for these synthetic PWID using the simulated HepCEP events including an indicator variable denoting if the individual would become infected in a clinical trial with a 1.5-year follow-up period and if the individual is HCV-susceptible (HCV RNA/Ab-negative).

2.1.2. Survival Analysis Models

To achieve its goals of minimizing sample size, PREDICTEE relies on a survival model that estimates the probability of HCV infection by the end of the clinical trial follow-up period, given the demographic, network, and injection characteristics of individual PWID. This trained survival model will be used in the PREDICTEE process to enrich the trial cohort with high-risk PWID most likely to experience acute HCV infection, increasing the cohort incidence of HCV and thereby decreasing the required sample size needed to attain adequate power. In our simulated HCV vaccine trials, we tested two models: (1) a Cox proportional hazards model, a classic survival analysis model [34], and (2) a non-parametric random survival forest (RSF), a more recently developed ensemble prediction method [35]. For each simulation, we implemented a 20/80 train-test split, in which 20% of the synthetic population of 123,071 PWID was used to train the survival analysis model, and the remaining 80% served as the recruitment pool from which simulated trial participants were recruited from. Both Cox and RSF models were trained using the demographic, behavioral, and network attributes provided by the HepCEP model, listed as follows:

1. Age;
2. Sex assigned at birth;
3. Syringe source (harm reduction program or other);
4. The number of PWID who gave the candidate drugs/injection equipment in the last 30 days;

5. The number of PWID who the candidate gave drugs/injection equipment to in the last 30 days;
6. The total number of people in the candidate's drug use network;
7. The number of daily injections;
8. The fraction of injections that involve receiving drugs or injection equipment from another person in the network.

The outcome variable used to train the models was the time (in years) until the PWID experience an acute HCV infection. Any PWID who did not experience an HCV infection during the 10-year HepCEP simulation were right-censored at 10 years. These trained Cox and RSF models are used to predict the probability that new PWID will be infected with HCV by 1.5 years after enrollment, allowing for the recruitment of high-risk PWID to maximize trial cohort incidence.

The RSF model also requires the selection of two main hyperparameters: number of trees (ntree) and number of variables randomly selected as candidates for splitting a node (mtry). Using the recommendations in the literature [36,37], we set $mtry = p/3$, with p being the number of explanatory variables; thus, $mtry = 3$ because there are eight explanatory variables used to train the model. Additionally, we set $ntree = 100$ based on [38,39].

2.1.3. Demographic Targets

While PREDICTEE primarily aims to achieve a minimal sample size through PWID risk prediction, it also attempts to improve trial generalizability by recruiting a trial cohort that resembles a prespecified target demographic. We balanced both objectives through a dynamic weighting and scoring process, which factors in both a candidate's HCV risk as well as the characteristics needed to reach the target demographic composition. Details on the scoring equation are described later in this section.

2.2. The PREDICTEE Workflow

An overview of PREDICTEE recruitment as it was simulated in this study is illustrated in Figure 1. PREDICTEE first requires pre-existing longitudinal data of the recruited population, containing characteristics predictive of HCV infection and a timeline of infection events. The synthetic population of 123,071 PWID serves this purpose for our simulations. We used 20% of these data to train a Cox or RSF model, and the remaining 80% served as the recruitment pool. PWID screening was simulated via random sampling of this recruitment pool. Screened candidates undergo immediate scoring, factoring in demographic considerations and their probability of HCV infection, estimated using the survival model. Once a batch of B candidates is screened and scored, the candidates with the highest scores receive HCV screening, and the top-scoring R HCV-susceptible PWID are recruited and randomized to vaccine or control groups.

After each batch, we update the weights used for scoring candidates based on the composition of the partial cohort. At a predetermined point, we also optionally perform sample size re-estimation based on the predicted incidence of the partial cohort, calculated by averaging the individual infection probabilities of the recruits, and allow for early termination of recruitment. The point at which sample size is re-estimated is inversely proportional to the estimated incidence improvement of the predictive model compared to conventional recruitment methods. For example, a model that is expected to double the incidence in the trial cohort would re-estimate sample size when half the required sample size of conventional recruitment is reached. These parameters would be determined via simulations using the same preliminary data that were used to train the predictive model. For a list of parameters used in PREDICTEE, see the Supplementary Materials. In subsequent sections, we describe the processes of weighting and scoring in greater detail.

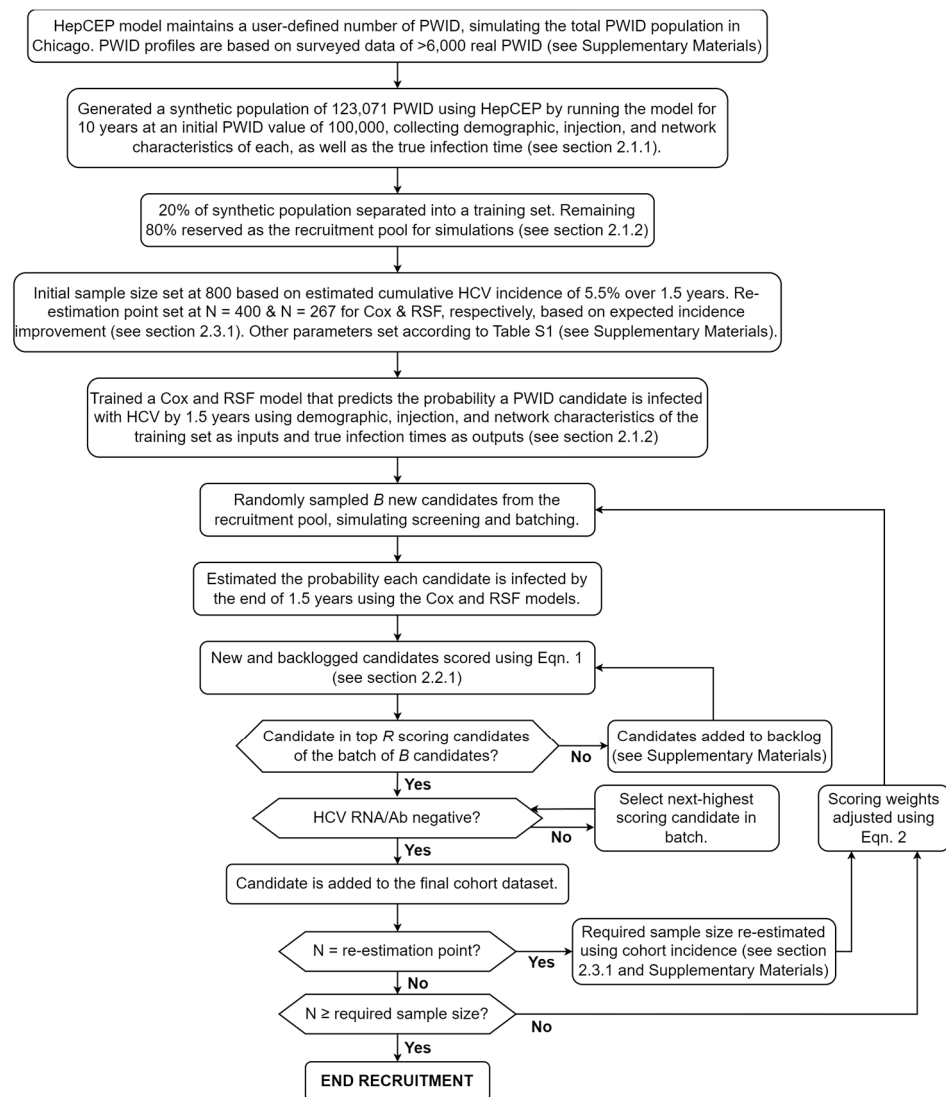


Figure 1. Schematic of the PREDICTEE recruitment method as it was performed in our simulations. During the setup, a predictive model is trained using existing longitudinal data, and parameters for PREDICTEE are set. Recruitment proceeds in cycles of scoring, batching, recruiting, and adjusting parameters. Optionally, sample size is re-estimated at a pre-specified number of PWID recruited (re-estimation point).

2.2.1. Candidate Scoring Equations

The arriving candidates are denoted c_1, c_2, \dots , and a new candidate c_i arrives at a time point t , where t represents the number of prior batches that have already been considered. The score $Score_t(c_i)$ is based on several factors: (1) the cohort of previously recruited subjects, denoted A_{t-1} , with individual recruited candidates denoted a ; (2) the target composition for the final cohort across each population category (e.g., proportion of non-Hispanic Black), denoted p_j , where $j \in J$, with J representing the larger set of all considered categories (demographic and/or others); and (3) the characteristics of candidate c_i . The latter includes the candidate's predicted incidence of HCV, denoted $I(c_i)$, as estimated by the predictive model, and the candidate's categories, denoted $d_j(c_i)$, which equal 1 when the person is a member of category j and 0 otherwise. The candidate score is given by Equation (1):

$$Score_t(c_i) = w_t I(c_i) + (100 - w_t) \sum_{j \in J} d_j(c_i) \left[p_j - \frac{1}{|A_{t-1}|} \sum_{a \in A_{t-1}} d_j(a) \right] \quad (1)$$

where w_t is the weight assigned to incidence and $100 - w_t$ is the weight of categorical factors. This scoring function simultaneously addresses the conflicting goals of trial equity while decreasing sample size via recruiting higher-risk candidates. For an example of candidate scoring, see Figure S1 in the Supplementary Materials.

After score assignment, the top- R -scoring HCV-naïve candidates are recruited. Because we test for susceptibility to HCV after scoring, if any of these R candidates are not HCV-susceptible, they are replaced by the next-best-scoring naïve candidate. A trial could optionally maintain a backlog of candidates who were not recruited and consider them alongside subsequent batches, which would help decrease batch variability and maintain a steady recruitment screening, particularly in situations when the value of B is small. This process is repeated until the end sample size is reached. The values of B and R can be altered based on trial capabilities. The ratio between them, R/B , is the proportion of each batch that is recruited. Decreasing the ratio minimizes the size of the sample, while increasing it minimizes the cost of screening.

Incidence and demographic representativeness are weighted dynamically, with PREDICTEE prioritizing incidence at the beginning of the trial and gradually giving more weight to closing population gaps as the trial advances. Thus, the initial w_0 is always 100, and then it decreases after each batch according to Equation (2):

$$w_t = w_{t-1} - \frac{R}{N} (w_{t-1} - w_{min}) \quad (2)$$

where w_t is the incidence weight for the current batch t , w_{t-1} is the incidence weight for the last batch $t-1$, w_{min} is the incidence weight floor, R is the number of candidates recruited per batch, and N is the most recent estimated sample size. For simulations where we anticipate a large gap from targets, we accelerated the rate of decrease, as described in the Supplementary Materials. This weighting approach would allow investigators to first establish a baseline profile of recruited candidates at the beginning of a trial based on the highest-risk candidates screened. As the trial progresses, this profile of candidates can be leveraged to address sampling inequities without significantly compromising the main objective of maintaining a low sample size.

Optionally, the trial plan could allow for early stopping of recruitment based on the cohort incidence. Namely, the stopping decision could consider the expected number of events in the recruited cohort as estimated by the model, without unblinding the arms or waiting to observe actual events. It is recommended that the decision threshold include a margin of safety for any model error.

2.3. Evaluating PREDICTEE HCV Vaccine Trial in Chicago PWID Population

2.3.1. Design of Simulation Experiments

In previous work, we simulated trials of HCV vaccines end-to-end [32,33,40,41]. In this work, we implement software that simulates only the outcome of the recruitment process. We applied PREDICTEE to two challenges in building a representative cohort: (a) a study cohort matching the Chicago area's PWID population in terms of race/ethnicity, sex, and age distribution; (b) a maximally balanced cohort with a racial composition target for Hispanic, non-Hispanic Black (NHBlack), non-Hispanic White (NHWhite), and Other individuals of 33:33:33:1, respectively, and a sex composition of 50:50 (male–female). For simulations in which age was targeted, each candidate was categorized into ten-year age groups. Age was not considered in the maximally balanced cohort (Scenario b) because of the extremely low prevalence (~3.5% of the recruitment pool) and HCV incidence in candidates in the 49+ age group in our Chicago PWID data. This would make it difficult to match a balanced cohort without arbitrarily decreasing the target for candidates in

the 49+ group. To compare PREDICTEE to conventional recruitment strategies, we also simulated two additional recruitment strategies using the Chicago area's PWID dataset, as follows. (1) Random/uniform: candidates are an unbiased random sample from the HCV-susceptible PWID population; (2) in-network: only candidates who receive syringes from their social network are recruited because they have a higher incidence rate (as proposed in [32,33,42]). Random recruitment is offered as a synthetic benchmark, although few trials can realistically implement the random recruitment of PWID due to the complexity of working with the PWID population, and it is avoided in sites like Chicago due to the low incidence rate.

For both Cox and RSF models, PREDICTEE was executed 100 times for each of 100 random train/test splits of the synthetic population, yielding 10,000 total trial recruitment simulations. For each simulation, we calculate Harrell's concordance index (C-index) for the unique trained model. The C-index is a goodness-of-fit measurement commonly used for survival analysis models with censored data, analogous to the area under the ROC curve (AUC) for more classic predictive models and diagnostic tests [43,44]. Based on the existing literature, the C-index for a survival analysis model should be at least 0.7 to adequately discriminate between risk profiles [45–47]. For all simulation runs, assumed vaccine efficacy was 60% and the trial follow-up period was 1.5 years, following the work of Page et al. [2]. We define vaccine efficacy as the proportionate reduction in chronic cases between the unvaccinated and vaccinated groups [48]. The initial required sample size was set to 800 to achieve 80% power based on the estimated 1.5-year cumulative incidence of 5.5% from the in-network recruitment method. Parameters were set accordingly: $B = 50$, $R = 5$, and $w_{min} = 25$. The candidate arrival process was simulated via random sampling from the recruitment pool. Simulations of random sampling and in-network recruitment were also repeated 10,000 times.

Sample size re-estimation occurred after 400 candidates recruited for Cox PREDICTEE and 267 candidates recruited for RSF PREDICTEE based on experimental data that showed a required sample size of approximately 800 candidates for in-network recruitment and that the Cox and RSF model would approximately double and triple the incidence, respectively. At this re-estimation point, the predicted incidence of the partial cohort is used to update the sample size rather than looking at the observed outcomes of recruits. This is achieved by averaging the infection probabilities generated by the Cox or RSF models for all candidates recruited to the trial up to the re-estimation point, without revealing treatment assignments. Due to the blinded nature of this process, there is a minimal impact on type I error [49]. Specific parameters and formulae involved in sample size re-estimation can be found in the Supplementary Materials.

Incidence values in this study represent cumulative incidence over the 1.5-year trial follow-up period, expressed as the proportion of recruited participants who develop HCV. Predicted incidence in PREDICTEE simulations is compared to observed incidence to confirm model validity (see the Supplementary Materials). Demographic statistics are recorded for each trial cohort and averaged across all runs for both Cox and RSF PREDICTEE. These are plotted alongside the target population and in-network recruitment to evaluate PREDICTEE's demographic adjustment capability.

For each recruitment method and demographic category, we also calculate participation-to-prevalence ratio (PPR), a metric that evaluates demographic representation and has been widely used in clinical trial settings for assessing adequate enrollment of demographic subgroups [50–54]. $PPR = \text{proportion among trial participants} / \text{proportion among disease population}$. Thus, a PPR of 1.0 for a specific demographic represents perfect representation. A PPR of less than 0.8 signifies underrepresentation, a PPR between 0.8 and 1.2 signifies adequate representation, and a PPR greater than 1.2 signifies overrepresentation. For our simulations, we report the average of the lowest PPR across all target categories in each run, denoted PPR_{min} .

2.3.2. Software Modeling Platform and Tools

Trial power analysis and simulations were conducted using R, with Cox and RSF models trained using the survival and randomForestSRC packages, respectively [55–57].

3. Results

Our analysis below compares recruitment techniques in terms of required sample size, representativeness, and screening requirements. We also compared PREDICTEE using the Cox model against RSF.

3.1. Matching the Chicago PWID Population

The baseline characteristics of the Chicago PWID population are listed in Table 1. Unless otherwise specified, PREDICTEE attempts to match the HCV-susceptible PWID population (third column)—this is the population most likely to eventually receive the vaccine. In-network recruitment represents a good method for recruitment into HCV trials without a predictive model and serves as a comparison for PREDICTEE. As can be seen in Table 1, in-network recruitment results in a significant underrepresentation of non-Hispanic Black candidates compared to the HCV-susceptible population.

Table 1. Demographic, behavioral, and network attributes of PWID populations in Chicago.

Attribute	2018 Chicago PWID Population	Susceptible Population with Receptive Network	HCV-Susceptible Population
<i>Demographic Attributes</i>			
Location (by ZIP Code)	City: 45.4% Suburbs: 54.6%	City: 27.5% Suburbs: 72.5%	City: 36.5% Suburbs: 63.5%
Race/Ethnicity	Hispanic: 18.7% NH Black: 20.8% NH White: 57.2% Other: 3.2%	Hispanic: 17.5% NH Black: 10.0% NH White: 69.3% Other: 3.2%	Hispanic: 18.1% NH Black: 15.5% NH White: 63.2% NH Other: 3.2%
Sex	Female: 30.6% Male: 69.4%	Female: 37.8% Male: 62.2%	Female: 31.5% Male: 68.5%
Age, mean (IQR)	35.2 (26.0–43.0)	29.8 (24.0–34.0)	31.4 (24.9–37.0)
Elapsed years of injection drug use, mean (IQR)	11.3 (3.4–15.6)	6.3 (2.0–8.6)	7.2 (2.5–9.9)
Enrollment in any harm-reduction (HR) program	HR: 48.4% Non-HR: 51.6%	HR: 33.1% Non-HR: 66.9%	HR: 45.5% Non-HR: 54.5%
HCV Infection State	Infected (acute or chronic): 28.1% Recovered (antibody-positive): 12.3%	0%—all susceptible	0%—all susceptible
<i>Behavioral Attributes</i>			
Daily Drug Injections, mean (IQR)	2.5 (0.9–3.3)	2.6 (0.9–3.6)	2.4 (0.8–3.2)
Probability of Receptible Sharing, mean (IQR)	19.4% (0.0–37.3%)	30.3% (5.0–50.0%)	21.3% (0.0–40.6%)
<i>Network Attributes</i>			
In Degree (Receptive Network Size)	0 (no network)—68.0% 1–23.5% ≥2–8.5%	0 (no network)—0% 1–75.7% ≥2–24.3%	0 (no network)—68.0% 1–23.5% ≥2–8.5%
Out Degree (Giving Network Size)	0 (no network)—72.0% 1–20.6% ≥2–7.4%	0 (no network)—45.0% 1–40.5% ≥2–14.5%	0 (no network)—69.4% 1–22.7% ≥2–7.9%

PREDICTEE led to a marked increase in cumulative incidence over the course of a simulated trial, as shown in Table 2a. Application of the Cox model led to a nearly two-fold increase in incidence compared to in-network recruitment from 0.055 (95%: 0.044–0.068) to 0.097 (95%: 0.090–0.104), and the RSF model led to a nearly three-fold increase to 0.149 (95%: 0.141–0.155). This corresponds to a sample size of 444 and 278 for Cox and RSF PREDICTEE, respectively, compared to 802 for in-network recruitment (a ratio of 1.81 and 2.88, respectively). In terms of screening requirements, Cox PREDICTEE achieved a smaller sample size with an approximate 12% increase in eligibility screening; however, RSF PREDICTEE achieved its reduction in sample size while also reducing screening requirements by almost 30%.

Table 2. Incidence data when PREDICTEE attempts to match two different demographic compositions: (a) Chicago’s susceptible PWID and (b) an arbitrary balanced population. All values represent the mean of 10,000 simulations, with a 95% range calculated using quantiles.

	Random Sample	In-Network Recruitment	PREDICTEE (Cox Model)	PREDICTEE (RSF Model)
(a) Matching Chicago’s Susceptible PWID				
Cohort Incidence	0.024 (0.018–0.033)	0.055 (0.044–0.068)	0.097 (0.090–0.104)	0.149 (0.141–0.155)
Required Sample Size (Calculated Using Cohort Incidence)	1876 (1356–2512)	802 (642–1010)	444 (408–480)	278 (264–294)
Expected Number of Candidates Screened Before Achieving Required Sample Size *	2207 (1595–2955)	4648 (3721–5853)	5224 (4800–5647)	3271 (3106–3459)
Post Hoc Power if 800 Recruited	49.2% (40.5–60.5%)	80.0% (71.5–87.1%)	95.5% (94.2–96.7%)	99.5% (99.3–99.6%)
PPR _{min} **	-	0.475 (0.356–0.568)	0.764 (0.593–0.934)	0.754 (0.685–0.834)
(b) Matching Arbitrary Balanced Demographics				
Cohort Incidence	0.024 (0.018–0.033)	0.055 (0.044–0.068)	0.085 (0.079–0.095)	0.137 (0.130–0.144)
Required Sample Size (Calculated Using Cohort Incidence)	1876 (1356–2512)	802 (642–1010)	506 (452–550)	304 (288–322)
Expected Number of Candidates Screened Before Achieving Required Sample Size *	2207 (1595–2955)	4648 (3721–5853)	5953 (5318–6471)	3576 (3388–3788)
Post Hoc Power if 800 Recruited	49.2% (40.5–60.5%)	80.0% (71.5–87.1%)	93.1% (91.3–95.2%)	99.2% (98.9–99.4%)
PPR _{min} ***	0.433 (0.396–0.472)	0.250 (0.221–0.280)	0.903 (0.889–0.918)	0.807 (0.792–0.821)

* Assuming a refusal rate of 15% [2], with 20.3% of PWID being both eligible for in-network recruitment and susceptible (calculated from HepCEP data) and 10% of PREDICTEE candidates being recruited. Values do not account for any additional inclusion/exclusion criteria not mentioned, and thus true screening numbers may be larger. ** Refers to the susceptible Chicago PWID population. PPR_{min} was not calculated for random sample because it is sampled directly from the target population of susceptible Chicago PWID. *** Refers to the arbitrary maximally balanced target population.

While improving incidence, PREDICTEE simultaneously corrected for deviations in demographic representativeness seen in in-network recruitment, illustrated by Figure 2A. This is also expressed in the PPR_{min} row in Table 2a, which shows that PPR_{min} increased

greatly from 0.475 (95%: 0.356–0.568) with in-network recruitment to 0.764 (95%: 0.593–0.934) for Cox and 0.754 (95%: 0.685–0.834) for RSF. It should be noted that PPR_{\min} remained under 0.80 for PREDICTEE recruitment exclusively because of the over-49 age group that is being matched. This is due to the low PWID prevalence and HCV incidence of this demographic, leading to continued underrepresentation. If this demographic were not considered in matching, PPR_{\min} was calculated to be 0.947 (95%: 0.924–0.970) for Cox PREDICTEE and 0.964 (95%: 0.948–0.974) for RSF PREDICTEE.

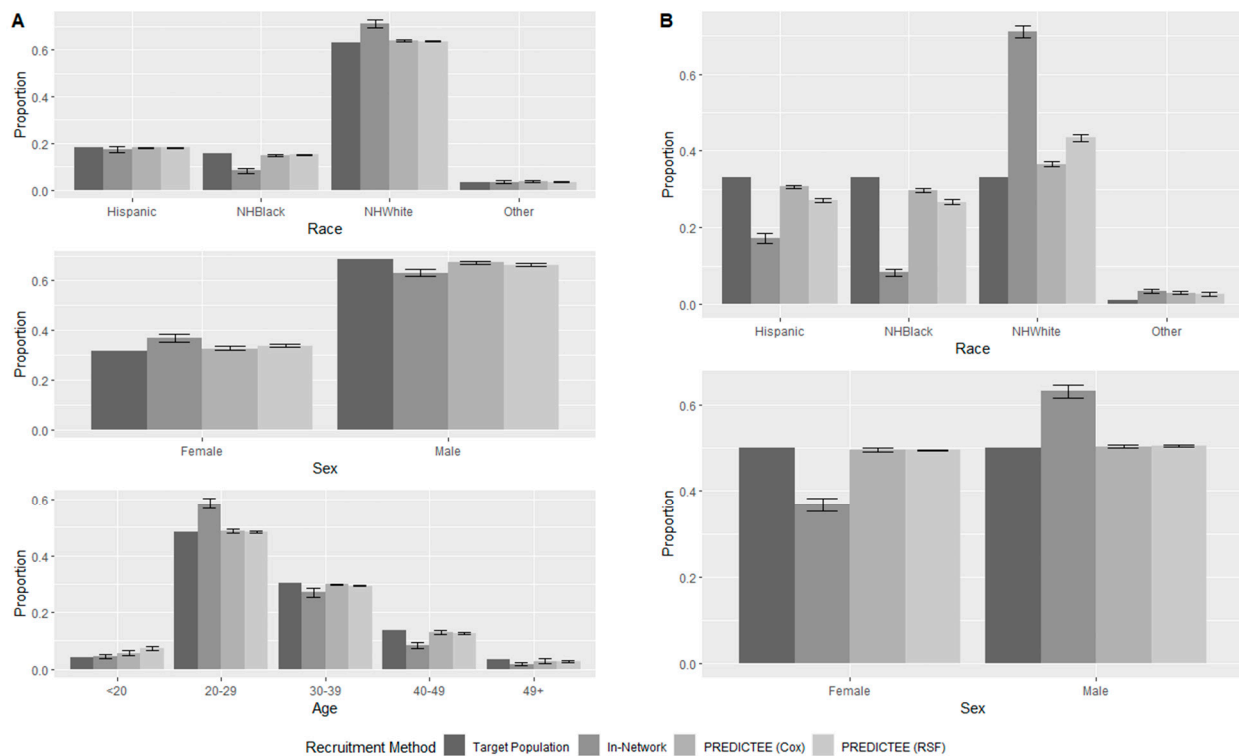


Figure 2. Demographic outcome of recruitment under two different target populations: (A) Chicago’s susceptible PWID and (B) an arbitrary balanced population. Error bars represent the 95% range over 10,000 trials. Sex reflects assignment at birth.

3.2. Targeting Arbitrary Demographics

Results for matching a maximally balanced population to assess the ability of PREDICTEE to adjust to more dissimilar targets are summarized in Table 2b and Figure 2B. In this scenario, in-network recruitment expressed significant inequities with a PPR_{\min} of 0.250 (95%: 0.221–0.280), while PREDICTEE was able to close demographic gaps reasonably well, with an average PPR_{\min} across the simulations of 0.903 (95%: 0.889–0.918) and 0.807 (95%: 0.792–0.821) for Cox and RSF PREDICTEE, respectively. At the same time, PREDICTEE increased cumulative incidence to 0.085 (95%: 0.079–0.095) and 0.137 (95%: 0.130–0.144), respectively. This corresponds to a required sample size of 506 and 304 for Cox and RSF PREDICTEE compared to 802 for in-network recruitment (a ratio of 1.58 and 2.64, respectively). Cox PREDICTEE achieved a smaller sample size with an approximately 28% increase in eligibility screening, while RSF PREDICTEE achieved its reduction in sample size while also reducing screening requirements by roughly 23% compared to in-network recruitment.

The models were generally quite accurate in predicting time to infection in the simulated epidemic; across 10,000 simulations, the average C-index value was 0.871 for Cox models and 0.946 for RSF models, indicating adequate risk profile discrimination.

4. Discussion

Our study developed PREDICTEE as a proposed method for reducing the required sample size within a vaccine clinical trial while ensuring cohort representativeness to a pre-specified target population. We have shown that contemporary techniques in recruitment may result in demographic differences within the trial cohort (Table 1), which may contribute to a lack of generalizability if these characteristics are relevant to vaccine efficacy or virus epidemiology. Through simulation, PREDICTEE illustrated its ability to decrease sample size two- to three-fold while also improving recruitment equity of target characteristics compared to conventional recruitment. It achieved this by selectively recruiting high-incidence PWID while prioritizing underrepresented demographics (Table 2a and Figure 2A). This was the case even in circumstances with a highly dissimilar target population (Table 2b and Figure 2B).

4.1. Implications

PREDICTEE complements and contributes to the literature on trial design strategies that leverage predictive models. Previous work examined ideas such as prognostic enrichment, adaptive designs, and sample size re-estimation. The statistical and clinical benefits of these strategies have been well documented, allowing trials to not only improve their efficiency and impact but also to ensure adequate statistical power [58,59]. We show in a simulated setting that PREDICTEE is able to maintain the benefits that adaptive trial designs offer while also providing further benefits to clinical trial recruitment through the implementation of both a predictive model and a weighting system prioritizing underrecruited categories of candidates.

A practical drawback of previous adaptive enrichment designs is that patient recruitment may be halted before interim analysis so that primary outcomes can be assessed, increasing trial duration and delaying submission of experimental agents for approval [60,61]. PREDICTEE presents a potential solution to this as it offers a predicted incidence value. Sample size could be re-estimated without unblinding, using only the data recorded at enrollment and before any events are observed. As a result, clinical trials would not need to pause recruitment. Naturally, this scheme relies on a well-calibrated predictive model and may necessitate a margin for any model error.

We also consider the ability of PREDICTEE to serve as an alternative to existing estimators that aim to generalize the results of clinical trials to target populations by accounting for nonrandom sampling [62,63]. By ensuring concordance between desired characteristics such as demographics between the trial cohort and the target population, PREDICTEE essentially aims to generalize trial results via equitable recruitment processes instead of post hoc quantitative methods that may be non-robust to model misspecification or the misestimation of parameters [64].

4.2. Limitations

Our quantitative results are based on agent-based simulations of a PWID population from a large metropolitan city. Although these simulations have been shown to be more realistic than aggregate models [65], they are still a simplification of human behavior and network formation. As a result, the survival models might be accurate in the simulation setting but experience increased prediction error in a real clinical trial environment. However, we believe that these simplifications are unlikely to affect the overall conclusions of this study, since simulations were only used to train the predictive models and evaluate the PREDICTEE workflow. The candidate data and recruitment pool that was used in these simulations were derived from real survey data collected from PWID in Chicago (see the Supplementary Materials). As in conventional trials, trial organizers may incorporate a margin of error into their sample size calculations to reduce the likelihood of an underpowered trial.

In designing our PREDICTEE recruitment simulations, we aimed to reflect real-world conditions, but some effects were not captured. Due to a lack of appropriate data on

the ease of recruitment and network referrals between PWID, we formed batches for PREDICTEE by randomly sampling from a predefined recruitment pool. Actual feeder processes for RCTs likely observe clustering and uneven sampling due to differences in accessibility of PWID [66,67], but we do not anticipate this having a tangible effect on outcomes. Additionally, loss to follow-up is higher in select subgroups (e.g., those engaged in higher risk activities) [68–70], but given the appropriate data, these effects could be easily incorporated into PREDICTEE. The PREDICTEE scheme also increases the HCV incidence in underrepresented demographics to a smaller extent than other demographics; however, this did not have a significant effect in our study (see the Supplementary Materials).

While our simulation results are based on metropolitan PWID data from the Chicago area, we anticipate that PREDICTEE will work in other geographic areas with similar HCV epidemics. However, each site has unique HCV risk factors, and ideally, users should train their own site-specific models. Similarly, all parameters of PREDICTEE (see the Supplementary Materials) should be designed based on local knowledge of the trial site. PREDICTEE would ideally rely on ample and accurate longitudinal local data of the target population to achieve optimal results. From a global lens, this study heavily incorporated a North American context through the use of ethno-racial categories that are prominent in the U.S.; however, PREDICTEE could be applied to a range of categories of interest such as genetic variants, socioeconomic categories, and more.

Finally, more research is needed on the cost-effectiveness of PREDICTEE in comparison to traditional recruitment methods. Although we anticipate that cost savings associated with recruiting fewer subjects would outweigh any additional screening costs with PREDICTEE, other costs would also have to be considered such as predictive model training, algorithm implementation, and personnel costs. This would require further studies on specific real-world PREDICTEE parameters and how they compare to traditional recruitment, such as total clinical trial duration, required volume of pre-trial data, and ease of identifying high-risk PWID.

4.3. Extensions

Future research with PREDICTEE should focus on refining the methodology in field conditions and validating its use in other types of geographical regions (e.g., rural). We discuss some possible real-world implementations of PREDICTEE in the Supplementary Materials. The PREDICTEE strategy could also incorporate extensions in which the arms within the trial are balanced both in terms of static characteristics, such as race and sex, as well as the number of predicted infectious events. Imbalance in post-randomization events [71] has been identified as an important source of trial failure in HCV vaccines [41], and it could be reduced by using the survival model to balance the trial arms. Efficiency of recruitment could also be enhanced by using predictive models beyond predicting incidence to, e.g., predict viral status, refusal probability, and/or nonadherence, among others, as demonstrated in a recent COVID-19 vaccine study [72]. The efficiency of predictive recruitment could benefit from schemes for dynamic model retraining, allowing trials to be carried out in sites with limited previous incidence data. Moreover, the adoption of PREDICTEE could be accelerated by the sharing of PWID and HCV epidemiological data via an application programming interface (API). While this study illustrated the benefits of PREDICTEE in the context of an HCV vaccine trial, this method could be easily extended to trials of other vaccines, such as COVID-19 or HIV, or trials for novel treatments that also often seek high-incidence candidates. Moreover, we envision that PREDICTEE could be adapted to trials with non-binary outcomes, such as weight, A1c, etc.

5. Conclusions

PREDICTEE is the first recruitment method implementing a predictive model that balances incidence and population representation with the goal of producing more equitable and feasible clinical trials. Our results illustrated that PREDICTEE can recruit high-incidence participants while adjusting recruitment to multiple target populations.

Further research is warranted to validate PREDICTEE in field conditions and using a variety of trial designs.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/healthcare12060644/s1>, Supplementary Materials containing Supplemental Methods (SM) and Supplemental Results (SR). References [22,23,32,33,41,49,73–88] are cited in the supplementary materials.

Author Contributions: Conceptualization, R.C. and A.G.; methodology, R.C., A.G., H.D., E.T., J.O. and M.E.M.-A.; software, R.C. and A.G.; validation, R.C.; formal analysis, R.C.; investigation, R.C.; resources, E.T., M.E.M.-A., K.P., J.O., B.B., H.D. and A.G.; data curation, B.B. and E.T.; writing—original draft preparation, all authors; writing—review and editing, all authors; visualization, R.C.; supervision, A.G. and H.D.; project administration, A.G. and H.D.; funding acquisition, E.T., M.E.M.-A., J.O., B.B., H.D. and A.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Institutes of Health (NIH) grant number R01-AI158666 “Computational modeling for HCV vaccine trial design and optimal vaccine-based combination interventions”. The NIH had no involvement in the conception of this study, data collection and analysis, interpretation of data, or in the decision to submit this manuscript for publication.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Our source code repository (<https://github.com/sashagutfraind/vaccinetrials> (accessed on 19 February 2024)) provides a copy of a spreadsheet alongside all analytical code. Publicly available synthetic data for the Chicago area can be found at <https://zenodo.org/record/21714#.YshzX-zMLmo> (accessed on 19 February 2024) [32]. Original survey data were licensed to the authors by third-party investigators. Readers may request access and approval will be considered on a case-by-case basis.

Acknowledgments: We thank Ilya Lipkovich and Marian Major for providing helpful comments on the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. The Lancet Gastroenterology & Hepatology. The Hunt for a Vaccine for Hepatitis C Virus Continues. *Lancet Gastroenterol. Hepatol.* **2021**, *6*, 253. [CrossRef] [PubMed]
2. Page, K.; Melia, M.T.; Veenhuis, R.T.; Winter, M.; Rousseau, K.E.; Massaccesi, G.; Osburn, W.O.; Forman, M.; Thomas, E.; Thornton, K.; et al. Randomized Trial of a Vaccine Regimen to Prevent Chronic HCV Infection. *N. Engl. J. Med.* **2021**, *384*, 541–549. [CrossRef] [PubMed]
3. Rosenberg, E.S.; Rosenthal, E.M.; Hall, E.W.; Barker, L.; Hofmeister, M.G.; Sullivan, P.S.; Dietz, P.; Mermin, J.; Ryerson, A.B. Prevalence of Hepatitis C Virus Infection in US States and the District of Columbia, 2013 to 2016. *JAMA Netw. Open* **2018**, *1*, e186371. [CrossRef]
4. Hofmeister, M.G.; Rosenthal, E.M.; Barker, L.K.; Rosenberg, E.S.; Barranco, M.A.; Hall, E.W.; Edlin, B.R.; Mermin, J.; Ward, J.W.; Ryerson, A.B. Estimating Prevalence of Hepatitis C Virus Infection in the United States, 2013–2016. *Hepatology* **2019**, *69*, 1020–1031. [CrossRef] [PubMed]
5. Shiffman, M.L. Hepatitis C Virus Therapy in the Direct Acting Antiviral Era. *Curr. Opin. Gastroenterol.* **2014**, *30*, 217–222. [CrossRef] [PubMed]
6. Bethea, E.D.; Chen, Q.; Hur, C.; Chung, R.T.; Chhatwal, J. Should We Treat Acute Hepatitis C? A Decision and Cost-Effectiveness Analysis. *Hepatology* **2018**, *67*, 837–846. [CrossRef] [PubMed]
7. Bruggmann, P.; Litwin, A.H. Models of Care for the Management of Hepatitis C Virus among People Who Inject Drugs: One Size Does Not Fit All. *Clin. Infect. Dis.* **2013**, *57* (Suppl. S2), S56–S61. [CrossRef]
8. Hellard, M.; Doyle, J.S.; Sacks-Davis, R.; Thompson, A.J.; McBryde, E. Eradication of Hepatitis C Infection: The Importance of Targeting People Who Inject Drugs. *Hepatology* **2014**, *59*, 366–369. [CrossRef]
9. Carlisle, B.; Kimmelman, J.; Ramsay, T.; MacKinnon, N. Unsuccessful Trial Accrual and Human Subjects Protections: An Empirical Analysis of Recently Closed Trials. *Clin. Trials* **2015**, *12*, 77–83. [CrossRef]
10. Huang, G.D.; Bull, J.; Johnston McKee, K.; Mahon, E.; Harper, B.; Roberts, J.N. Clinical Trials Recruitment Planning: A Proposed Framework from the Clinical Trials Transformation Initiative. *Contemp. Clin. Trials* **2018**, *66*, 74–79. [CrossRef]

11. Goldberg, A.; Bakhireva, L.N.; Page, K.; Henrie, A.M. A Qualitative Scoping Review of Early-Terminated Clinical Trials Sponsored by the Department of Veterans Affairs Cooperative Studies Program From 2010 to 2020. *Epidemiol. Rev.* **2022**, *44*, 110–120. [[CrossRef](#)]
12. Matheny, J.G. The Economics of Pharmaceutical Development: Costs, Risks, and Incentives. Ph.D. Thesis, Johns Hopkins University, Baltimore, MD, USA, 2013.
13. Battelle Technology Partnership Practice. *Biopharmaceutical Industry-Sponsored Clinical Trials: Impact on State Economies*; Battelle: Columbus, OH, USA, 2015.
14. Wilder, J.; Saraswathula, A.; Hasselblad, V.; Muir, A. A Systematic Review of Race and Ethnicity in Hepatitis C Clinical Trial Enrollment. *J. Natl. Med. Assoc.* **2016**, *108*, 24–29. [[CrossRef](#)]
15. Grebely, J.; Raffa, J.D.; Lai, C.; Krajden, M.; Conway, B.; Tyndall, M.W. Factors Associated with Spontaneous Clearance of Hepatitis C Virus among Illicit Drug Users. *Can. J. Gastroenterol.* **2007**, *21*, 447–451. [[CrossRef](#)]
16. van den Berg, C.H.B.S.; Grady, B.P.X.; Schinkel, J.; van de Laar, T.; Molenkamp, R.; van Houdt, R.; Coutinho, R.A.; van Baarle, D.; Prins, M. Female Sex and IL28B, a Synergism for Spontaneous Viral Clearance in Hepatitis C Virus (HCV) Seroconverters from a Community-Based Cohort. *PLoS ONE* **2011**, *6*, e27555. [[CrossRef](#)]
17. Bakr, I.; Rekecewicz, C.; Hosseiny, M.E.; Ismail, S.; Daly, M.E.; El-Kafrawy, S.; Esmat, G.; Hamid, M.A.; Mohamed, M.K.; Fontanet, A. Higher Clearance of Hepatitis C Virus Infection in Females Compared with Males. *Gut* **2006**, *55*, 1183–1187. [[CrossRef](#)] [[PubMed](#)]
18. Piasecki, B.A.; Lewis, J.D.; Reddy, K.R.; Bellamy, S.L.; Porter, S.B.; Weinrieb, R.M.; Stieritz, D.D.; Chang, K.-M. Influence of Alcohol Use, Race, and Viral Coinfections on Spontaneous HCV Clearance in a US Veteran Population. *Hepatology* **2004**, *40*, 892–899. [[CrossRef](#)] [[PubMed](#)]
19. Reid, M.; Price, J.C.; Tien, P.C. Hepatitis C Virus Infection in the Older Patient. *Infect. Dis. Clin. North Am.* **2017**, *31*, 827–838. [[CrossRef](#)] [[PubMed](#)]
20. FDA Office of Minority Health and Health Equity. Enhance EQUITY in Clinical Trials. Available online: <https://www.fda.gov/consumers/enhance-equity-initiative/enhance-equity-clinical-trials> (accessed on 27 July 2022).
21. National Institutes of Health. Inclusion of Women and Minorities as Participants in Research Involving Human Subjects. Available online: <https://grants.nih.gov/policy/inclusion/women-and-minorities.htm> (accessed on 27 July 2022).
22. Kerr, K.F.; Roth, J.; Zhu, K.; Thiessen-Philbrook, H.; Meisner, A.; Wilson, F.P.; Coca, S.; Parikh, C.R. Evaluating Biomarkers for Prognostic Enrichment of Clinical Trials. *Clin. Trials* **2017**, *14*, 629–638. [[CrossRef](#)]
23. Irazabal, M.V.; Abebe, K.Z.; Bae, K.T.; Perrone, R.D.; Chapman, A.B.; Schrier, R.W.; Yu, A.S.; Braun, W.E.; Steinman, T.I.; Harris, P.C.; et al. Prognostic Enrichment Design in Clinical Trials for Autosomal Dominant Polycystic Kidney Disease: The HALT-PKD Clinical Trial. *Nephrol. Dial. Transplant.* **2017**, *32*, 1857–1865. [[CrossRef](#)] [[PubMed](#)]
24. Heitjan, D.F.; Ge, Z.; Ying, G. Real-Time Prediction of Clinical Trial Enrollment and Event Counts: A Review. *Contemp. Clin. Trials* **2015**, *45*, 26–33. [[CrossRef](#)] [[PubMed](#)]
25. Stevens, V.W.; Russo, E.M.; Young-Xu, Y.; Leecaster, M.; Zhang, Y.; Zhang, C.; Yu, H.; Cai, B.; Gonzalez, E.N.; Gerding, D.N.; et al. Identification of Patients at Risk of Clostridioides Difficile Infection for Enrollment in Vaccine Clinical Trials. *Vaccine* **2021**, *39*, 536–544. [[CrossRef](#)] [[PubMed](#)]
26. Kahn, J.M.; Gray, D.M., II; Oliveri, J.M.; Washington, C.M.; DeGraffinreid, C.R.; Paskett, E.D. Strategies to Improve Diversity, Equity, and Inclusion in Clinical Trials. *Cancer* **2022**, *128*, 216–221. [[CrossRef](#)]
27. Kang, M.; Nicolay, U. Evaluation of Operational Chronic Infection Endpoints for HCV Vaccine Trials. *Contemp. Clin. Trials* **2008**, *29*, 671–678. [[CrossRef](#)]
28. Young, A.M.; Stephens, D.B.; Khaleel, H.A.; Havens, J.R. Hepatitis C Vaccine Clinical Trials among People Who Use Drugs: Potential for Participation and Involvement in Recruitment. *Contemp. Clin. Trials* **2015**, *41*, 9–16. [[CrossRef](#)] [[PubMed](#)]
29. Ismail, A.; Al-Zoubi, T.; El Naqa, I.; Saeed, H. The Role of Artificial Intelligence in Hastening Time to Recruitment in Clinical Trials. *BJR Open* **2023**, *5*, 20220023. [[CrossRef](#)] [[PubMed](#)]
30. Harrer, S.; Shah, P.; Antony, B.; Hu, J. Artificial Intelligence for Clinical Trial Design. *Trends Pharmacol. Sci.* **2019**, *40*, 577–591. [[CrossRef](#)]
31. Oikonomou, E.K.; Thangaraj, P.M.; Bhatt, D.L.; Ross, J.S.; Young, L.H.; Krumholz, H.M.; Suchard, M.A.; Khera, R. An Explainable Machine Learning-Based Phenomapping Strategy for Adaptive Predictive Enrichment in Randomized Clinical Trials. *npj Digit. Med.* **2023**, *6*, 217. [[CrossRef](#)]
32. Gutfraind, A.; Boodram, B.; Prachand, N.; Hailegiorgis, A.; Dahari, H.; Major, M.E. Agent-Based Model Forecasts Aging of the Population of People Who Inject Drugs in Metropolitan Chicago and Changing Prevalence of Hepatitis C Infections. *PLoS ONE* **2015**, *10*, e0137993. [[CrossRef](#)]
33. Tatara, E.; Collier, N.T.; Ozik, J.; Gutfraind, A.; Cotler, S.J.; Dahari, H.; Major, M.; Boodram, B. Multi-Objective Model Exploration of Hepatitis C Elimination in an Agent-Based Model of People Who Inject Drugs. *Proc. Winter Simul. Conf.* **2019**, *2019*, 1008–1019. [[CrossRef](#)]
34. Cox, D.; Oakes, D. *Analysis of Survival Data*; CRC Press: Boca Raton, FL, USA, 1984; ISBN 978-0-412-24490-2.
35. Ishwaran, H.; Kogalur, U.B.; Blackstone, E.H.; Lauer, M.S. Random Survival Forests. *Ann. Appl. Stat.* **2008**, *2*, 841–860. [[CrossRef](#)]
36. Genuer, R.; Poggi, J.-M.; Tuleau, C. Random Forests: Some Methodological Insights. *arXiv* **2008**, arXiv:0811.3619.

37. Goldstein, B.A.; Polley, E.C.; Briggs, F.B.S. Random Forests for Genetic Association Studies. *Stat. Appl. Genet. Mol. Biol.* **2011**, *10*, 32. [CrossRef]
38. Oshiro, T.M.; Perez, P.S.; Baranauskas, J.A. How Many Trees in a Random Forest? In *Machine Learning and Data Mining in Pattern Recognition—Proceedings of the 8th International Conference, MLDM 2012, Berlin, Germany, 13–20 July 2012*; Perner, P., Ed.; Springer: Berlin/Heidelberg, 2012; pp. 154–168.
39. Probst, P.; Boulesteix, A.-L. To Tune or Not to Tune the Number of Trees in Random Forest? *J. Mach. Learn. Res.* **2017**, *18*, 1–18.
40. Tatara, E.; Gutfraind, A.; Collier, N.T.; Echevarria, D.; Cotler, S.J.; Major, M.E.; Ozik, J.; Dahari, H.; Boodram, B. Modeling Hepatitis C Micro-Elimination among People Who Inject Drugs with Direct-Acting Antivirals in Metropolitan Chicago. *PLoS ONE* **2022**, *17*, e0264983. [CrossRef]
41. Mackesy-Amiti, M.E.; Gutfraind, A.; Tatara, E.R.; Collier, N.T.; Cotler, S.J.; Page, K.; Ozik, J.T.; Boodram, B.; Major, M.E.; Dahari, H. Simulations of HCV Vaccine Trials Demonstrate Effects of Background Incidence and Unbalanced Exposure That Can Impact Vaccine Efficacy. *Hepatology* **2021**, *74*, 604A.
42. Gutfraind, A.; Mackesy-Amiti, M.E.; Tatara, E.R.; Collier, N.T.; Cotler, S.J.; Page, K.; Ozik, J.T.; Boodram, B.; Major, M.E.; Dahari, H. Simulations of HCV Vaccine Trials Reveal Opportunities to Re-Evaluate Vaccine Efficacy. *J. Hepatol.* **2021**, *75*, S768.
43. Harrell, F.E., Jr.; Califf, R.M.; Pryor, D.B.; Lee, K.L.; Rosati, R.A. Evaluating the Yield of Medical Tests. *JAMA* **1982**, *247*, 2543–2546. [CrossRef]
44. Uno, H.; Cai, T.; Pencina, M.J.; D’Agostino, R.B.; Wei, L.J. On the C-Statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. *Stat. Med.* **2011**, *30*, 1105–1117. [CrossRef]
45. Hartman, N.; Kim, S.; He, K.; Kalbfleisch, J.D. Pitfalls of the Concordance Index for Survival Outcomes. *Stat. Med.* **2023**, *42*, 2179–2190. [CrossRef]
46. Steyerberg, E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*; Springer: New York, NY, USA, 2009.
47. DeMaris, A.; Selman, S. *Converting Data into Evidence: A Statistics Primer for the Medical Practitioner*; Springer: New York, NY, USA, 2013.
48. Centers for Disease Control and Prevention Principles of Epidemiology. Available online: <https://www.cdc.gov/csels/dsepd/ss1978/lesson3/section6.html> (accessed on 1 July 2023).
49. Friede, T.; Pohlmann, H.; Schmidli, H. Blinded sample size reestimation in event-driven clinical trials: Methods and an application in multiple sclerosis. *Pharm. Stat.* **2019**, *18*, 351–365. [CrossRef] [PubMed]
50. Chen, S.; Li, J.; Shu, M. Use of Participation to Prevalence Ratio for Evaluating the Representation Status of Women in Oncology Clinical Trials. *JAMA Oncol.* **2022**, *8*, 479–480. [CrossRef]
51. Khan, S.U.; Khan, M.Z.; Raghu Subramanian, C.; Riaz, H.; Khan, M.U.; Lone, A.N.; Khan, M.S.; Benson, E.-M.; Alkhouli, M.; Blaha, M.J.; et al. Participation of Women and Older Participants in Randomized Clinical Trials of Lipid-Lowering Therapies: A Systematic Review. *JAMA Netw. Open* **2020**, *3*, e205202. [CrossRef]
52. Saltzman, R.G.; Jayaweera, D.T.; Caceres, L.V.; Tovar, J.A.; Vidro-Casiano, M.; Karakeshishyan, V.; Soto, J.; Khan, A.; Mitrani, R.D.; Schulman, I.H.; et al. Demographic Representation in Clinical Trials for Cell-Based Therapy. *Contemp. Clin. Trials Commun.* **2021**, *21*, 100702. [CrossRef]
53. Scott, P.E.; Unger, E.F.; Jenkins, M.R.; Southworth, M.R.; McDowell, T.-Y.; Geller, R.J.; Elahi, M.; Temple, R.J.; Woodcock, J. Participation of Women in Clinical Trials Supporting FDA Approval of Cardiovascular Drugs. *J. Am. Coll. Cardiol.* **2018**, *71*, 1960–1969. [CrossRef]
54. Jin, X.; Chandramouli, C.; Allocco, B.; Gong, E.; Lam, C.S.P.; Yan, L.L. Women’s Participation in Cardiovascular Clinical Trials From 2010 to 2017. *Circulation* **2020**, *141*, 540–548. [CrossRef] [PubMed]
55. Therneau, T.M.; Lumley, T.; Atkinson, E.; Crowson, C. A Package for Survival Analysis in R. Available online: <https://CRAN.R-project.org/package=survival> (accessed on 20 July 2022).
56. Ishwaran, H.; Kogalur, U.B. Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). Available online: <https://cran.r-project.org/web/packages/randomForestSRC/index.html> (accessed on 20 July 2022).
57. R Core Team R. *A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2019.
58. Simon, N.; Simon, R. Adaptive Enrichment Designs for Clinical Trials. *Biostatistics* **2013**, *14*, 613–625. [CrossRef] [PubMed]
59. Mehta, C.R.; Pocock, S.J. Adaptive Increase in Sample Size When Interim Results Are Promising: A Practical Guide with Examples. *Stat. Med.* **2011**, *30*, 3267–3284. [CrossRef]
60. Uozumi, R.; Yada, S.; Kawaguchi, A. Patient Recruitment Strategies for Adaptive Enrichment Designs with Time-to-Event Endpoints. *BMC Med. Res. Methodol.* **2019**, *19*, 159. [CrossRef]
61. Rosenblum, M.; Hanley, D.F. Adaptive Enrichment Designs for Stroke Clinical Trials. *Stroke* **2017**, *48*, 2021–2025. [CrossRef] [PubMed]
62. Buchanan, A.L.; Hudgens, M.G.; Cole, S.R.; Mollan, K.R.; Sax, P.E.; Daar, E.S.; Adimora, A.A.; Eron, J.J.; Mugavero, M.J. Generalizing Evidence from Randomized Trials Using Inverse Probability of Sampling Weights. *J. R. Stat. Soc. Ser. A Stat. Soc.* **2018**, *181*, 1193–1209. [CrossRef] [PubMed]
63. Li, F.; Buchanan, A.L.; Cole, S.R. Generalizing Trial Evidence to Target Populations in Non-Nested Designs: Applications to AIDS Clinical Trials. *J. R. Stat. Soc. Ser. C Appl. Stat.* **2022**, *71*, 669–697. [CrossRef]
64. Mansournia, M.A.; Altman, D.G. Inverse Probability Weighting. *BMJ* **2016**, *352*, i189. [CrossRef]

65. Kasereka, S.K.; Zohinga, G.N.; Kiketa, V.M.; Ngoie, R.-B.M.; Mputu, E.K.; Kasoro, N.M.; Kyandoghere, K. Equation-Based Modeling vs. Agent-Based Modeling with Applications to the Spread of COVID-19 Outbreak. *Mathematics* **2023**, *11*, 253. [[CrossRef](#)]
66. Paquette, D.; Bryant, J.; de Wit, J. Respondent-Driven Sampling and the Recruitment of People with Small Injecting Networks. *AIDS Behav.* **2012**, *16*, 890–899. [[CrossRef](#)]
67. Harris, M.; Rhodes, T. Hepatitis C Treatment Access and Uptake for People Who Inject Drugs: A Review Mapping the Role of Social Factors. *Harm Reduct. J.* **2013**, *10*, 7. [[CrossRef](#)] [[PubMed](#)]
68. Soria, J.; Johnson, T.; Collins, J.; Corby-Lee, G.; Thacker, J.; White, C.; Hoven, A.; Thornton, A. Risk Factors for Loss to Follow-up of Persons Who Inject Drugs Enrolled at Syringe Services Programs in Kentucky. *Int. J. Drug Policy* **2021**, *95*, 103255. [[CrossRef](#)] [[PubMed](#)]
69. Darvishian, M.; Wong, S.; Binka, M.; Yu, A.; Ramji, A.; Yoshida, E.M.; Wong, J.; Rossi, C.; Butt, Z.A.; Bartlett, S.; et al. Loss to Follow-up: A Significant Barrier in the Treatment Cascade with Direct-Acting Therapies. *J. Viral Hepat.* **2020**, *27*, 243–260. [[CrossRef](#)]
70. Levy, V.; Evans, J.L.; Stein, E.S.; Davidson, P.J.; Lum, P.J.; Hahn, J.A.; Page, K. Are Young Injection Drug Users Ready and Willing to Participate in Preventive HCV Vaccine Trials? *Vaccine* **2010**, *28*, 5947–5951. [[CrossRef](#)]
71. Gilbert, P.B.; Bosch, R.J.; Hudgens, M.G. Sensitivity Analysis for the Assessment of Causal Vaccine Effects on Viral Load in HIV Vaccine Trials. *Biometrics* **2003**, *59*, 531–541. [[CrossRef](#)]
72. Mewhirter, J.; Sagir, M.; Sanders, R. Towards a Predictive Model of COVID-19 Vaccine Hesitancy among American Adults. *Vaccine* **2022**, *40*, 1783–1789. [[CrossRef](#)]
73. Tempalski, B.; Pouget, E.R.; Cleland, C.M.; Brady, J.E.; Cooper, H.L.F.; Hall, H.I.; Lansky, A.; West, B.S.; Friedman, S.R. Trends in the Population Prevalence of People Who Inject Drugs in US Metropolitan Areas 1992–2007. *PLoS ONE* **2013**, *8*, e64789. [[CrossRef](#)]
74. Lansky, A.; Abdul-Quader, A.S.; Cribbin, M.; Hall, T.; Finlayson, T.J.; Garfein, R.S.; Lin, L.S.; Sullivan, P.S. Developing an HIV Behavioral Surveillance System for Injecting Drug Users: The National HIV Behavioral Surveillance System. *Public Health Rep.* **2007**, *122*, 48–55. [[CrossRef](#)]
75. Huo, D.; Ouellet, L.J. Needle Exchange and Injection-Related Risk Behaviors in Chicago: A Longitudinal Study. *JAIDS J. Acquir. Immune Defic. Syndr.* **2007**, *45*, 108–114. [[CrossRef](#)]
76. Boodram, B.; Hotton, A.L.; Shekhtman, L.; Gutfraind, A.; Dahari, H. High-Risk Geographic Mobility Patterns among Young Urban and Suburban Persons Who Inject Drugs and Their Injection Network Members. *J. Urban Health* **2018**, *95*, 71–82. [[CrossRef](#)]
77. Boodram, B.; Mackesy-Amiti, M.-E.; Latkin, C. The Role of Social Networks and Geography on Risky Injection Behaviors of Young Persons Who Inject Drugs. *Drug Alcohol Depend.* **2015**, *154*, 229–235. [[CrossRef](#)]
78. Temple, R. Enrichment of Clinical Study Populations. *Clin. Pharmacol. Ther.* **2010**, *88*, 774–778. [[CrossRef](#)]
79. Ferguson, T.S. Who Solved the Secretary Problem? *Stat. Sci.* **1989**, *4*, 282–289. [[CrossRef](#)]
80. Camidge, D.R.; Park, H.; Smoyer, K.E.; Jacobs, I.; Lee, L.J.; Askerova, Z.; McGinnis, J.; Zakharia, Y. Race and Ethnicity Representation in Clinical Trials: Findings from a Literature Review of Phase I Oncology Trials. *Future Oncol.* **2021**, *17*, 3271–3280. [[CrossRef](#)]
81. Ma, M.A.; Gutiérrez, D.E.; Frausto, J.M.; Al-Delaimy, W.K. Minority Representation in Clinical Trials in the United States: Trends Over the Past 25 Years. *Mayo Clin. Proc.* **2021**, *96*, 264–266. [[CrossRef](#)] [[PubMed](#)]
82. Kennedy-Martin, T.; Curtis, S.; Faries, D.; Robinson, S.; Johnston, J. A Literature Review on the Representativeness of Randomized Controlled Trial Samples and Implications for the External Validity of Trial Results. *Trials* **2015**, *16*, 495. [[CrossRef](#)]
83. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1988.
84. Stark, M.; Hesse, M.; Brannath, W.; Zapf, A. Blinded Sample Size Re-Estimation in a Comparative Diagnostic Accuracy Study. *BMC Med. Res. Methodol.* **2022**, *22*, 115. [[CrossRef](#)] [[PubMed](#)]
85. Venn, M.L.; Knowles, C.H.; Li, E.; Glasbey, J.; Morton, D.G.; Hooper, R. ESCP EAGLE Safe Anastomosis Collaborative Implementation of a Batched Stepped Wedge Trial Evaluating a Quality Improvement Intervention for Surgical Teams to Reduce Anastomotic Leak after Right Colectomy. *Trials* **2023**, *24*, 329. [[CrossRef](#)] [[PubMed](#)]
86. Yao, X.; Attia, Z.I.; Behnken, E.M.; Walvatne, K.; Giblon, R.E.; Liu, S.; Siontis, K.C.; Gersh, B.J.; Graff-Radford, J.; Rabinstein, A.A.; et al. Batch Enrollment for an Artificial Intelligence-Guided Intervention to Lower Neurologic Events in Patients with Undiagnosed Atrial Fibrillation: Rationale and Design of a Digital Clinical Trial. *Am. Heart J.* **2021**, *239*, 73–79. [[CrossRef](#)] [[PubMed](#)]
87. Wermuth, P. Participant Recruitment, Screening, and Enrollment. In *Principles and Practice of Clinical Trials*; Piantadosi, S., Meinert, C.L., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 257–278. ISBN 978-3-319-52636-2.
88. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.