Critical Review

Overcoming the "Valley of Death" in Medications Development for Alcohol Use Disorder

Lara A. Ray [iD], Spencer Bujarski, Daniel James Olan Roche [iD], and Molly Magill

As the development of novel pharmacotherapies for alcohol use disorder (AUD) has been slow, the discovery and testing of more efficacious pharmacotherapies for AUD represent a high priority research area. In fact, the transition from preclinical to clinical testing of novel compounds has been termed the "valley of death" in medications development. One key obstacle consists of the lack of an articulated set of goals for each stage of medications development. Specifically, the knowledge outputs required to make the transition from safety testing, to early efficacy detection, to confirming clinical efficacy remain unclear, and this is despite a great deal of interest and substantial financial investment in developing novel therapeutics for AUD. This qualitative critical review seeks to draw parallels and lessons from the well-established stage model for behavioral therapies research with alcohol and other substance use disorders and to apply these insights into AUD pharmacotherapy development. We argue that human laboratory models and/or pilot randomized controlled trials should serve as intermediaries in the transition from preclinical studies to large, and costly, randomized controlled efficacy trials. The relative strengths and weaknesses of pilot clinical trials versus human laboratory studies for bridging the "valley of death" are discussed and explored via a Monte Carlo data simulation study. Multiple permutations of suitable research designs informed by the behavioral therapies development model are discussed with the overall goal of promoting consilience and maximizing efficiency across all phases of clinical testing of novel AUD pharmacotherapies.

Key Words: Medications Development, Pharmacotherapy, Behavioral Therapies Research, Alcohol Use Disorder, Treatment.

P HARMACOTHERAPIES FOR ALCOHOL use disorder (AUD) are used less often than behavioral therapies (Fuller and Hiller-Sturmhofel, 1999). The limited use of pharmacotherapy for AUD is due, in part, to a very limited number of approved options. Currently, the only pharmacotherapies approved by the Food and Drug Administration (FDA) for the treatment of AUD are disulfiram (Antabuse®), acamprosate, oral naltrexone, and Vivitrol (i.e., an injectable extended-release formulation of naltrexone). Given the dearth of available pharmacotherapies, as well as their limited-to-moderate efficacy (Blodgett et al., 2014; Donoghue et al., 2015; Jonas et al., 2014; Rösner et al., 2010), medications development to treat AUD remains a top research priority (Litten et al., 2012a, 2016). In fact, current foci in medications development emphasize the identification of new molecular targets and the development of novel compounds for these targets (Litten et al., 2012a; Ray et al., 2014). This is consistent with the phenotypic complexity of AUD, which calls for multiple molecular targets that can effectively address multiple pathways of risk (Litten et al., 2015). Despite rigorous efforts in medications development for AUD over the past 2 decades, it is apparent that there are numerous obstacles to an efficient process of screening and developing novel compounds.

One primary obstacle is the transition from preclinical to clinical testing, which has been termed the "valley of death" in medications development (Litten et al., 2012a). Specifically, the process of shepherding novel compounds from promising preclinical results in animal models through Investigational New Drug (IND) application requirements and human testing has been historically fraught; several promising compounds never advanced from preclinical to human subjects testing due to a variety of issues. A number of structural and regulatory issues are at play and NIH-level initiatives have recently been put in place to facilitate this transition, including incentives for industry–academia collaborations (Litten et al., 2016). Although outside of the scope of the present review, excellent reviews and commentaries from National Institute on Alcohol Abuse and Alcoholism (NIAAA) scientists have outlined the necessity of industry–academia collaborations in medication development for AUD (Litten et al., 2012a, 2014a,b, 2016).

One of the largest issues contributing to the "valley of death" is the lack of infrastructure to support the transition from animal to human testing in academia. Further, even if a compound makes it past this valley, most researchers lack the financial resources, regulatory expertise, and marketing resources to move the medication through costly clinical trials and into the postapproval market. Partnerships with industry remove, or at least mitigate, many of these issues, and for the foreseeable future, it is highly unlikely that any medication will receive FDA approval without collaboration with pharmaceutical companies. However, even with industry support, a remaining barrier is a minimally articulated set of goals and methods for each stage of human testing. Specifically, the requirements to successfully transition from safety testing in humans, to early efficacy detection in clinical samples, and finally to confirming clinical efficacy remain unclear, despite a great deal of interest and substantial investments in medications development for AUD. This review seeks to draw parallels and lessons from the well-established stage model of behavioral therapies research for alcohol and drug abuse and to discuss its application to research on pharmacological treatment development. In this study, we provide theoretical and empirical support for the standard medication development practices that human laboratory models and/or pilot randomized controlled trials (RCTs) should serve as intermediaries in the transition from preclinical studies to large, and costly, randomized controlled efficacy trials. The relative strengths and weaknesses of pilot clinical trials versus human laboratory studies for bridging the "valley of death" are discussed and explored via a Monte Carlo data simulation study. The overarching goal of this review was to shed light on the optimal approach to efficiently and reliably translate preclinical findings on promising medications for AUD to human clinical populations suffering from the disorder.

STAGES OF MEDICATIONS DEVELOPMENT

In an influential article setting an agenda for medications development for AUD, Litten and colleagues (2012a) outlined a set of objectives for the next decade of research. They describe stages for both preclinical testing and clinical testing in drug development. Litten and colleagues (2012a) particularly emphasize the "valley of death" where the majority of medications that show promise in animal models never make the transition to human subjects testing. This obstacle to medications development persists despite many well-validated methods for preclinical testing of AUD pharmacotherapies (Egli, 2005). Next, we address why medications fail to be shepherded across this valley and propose potential solutions informed by behavioral therapies research.

Clinical testing for a pharmacological treatment consists of multiple stages. Phase 1 is comprised of safety, pharmacokinetics (pK), dose selection, and alcohol interaction studies in non–treatment-seeking samples. This early stage of human testing is sometimes broken down into Phase 1a and

1b, where Phase 1a focuses on safety, pK, and alcohol interaction studies, and Phase 1b tests initial efficacy in the human laboratory with either single or multiple medication doses. Phase 2 is comprised of early efficacy studies in the human laboratory as well as safety and dose finding. Like Phase 1, Phase 2 efficacy testing can be described as early Phase 2 (referring to human laboratory testing in non–treatment-seeking samples) and late Phase 2 (referring to initial RCTs in treatment-seeking samples). Notably, Phase 1b and early Phase 2 overlap considerably in terms of goals and methods. Late Phase 2 is specifically defined as an efficacy determination phase through initial RCTs with treatment-seeking individuals. Phase 3 consists of verifying efficacy and safety through placebo-controlled RCTs with treatment-seeking individuals and possibly through large-scale multisite trials that support a new drug application to the FDA. Finally, Phase 4 is defined by monitoring safety, implementation, and adoption of a novel treatment after FDA approval during postmarket safety monitoring. Regardless of the preferred nomenclature for the stage model of medications development, there is a general recognition that clinical testing should progress from first determining safety, to confirmation of both clinical efficacy and safety, to subsequent regulatory approval.

While the clinical trial stages of medication development are fairly well defined, the outcomes needed to advance a medication through early phases are ambiguous, and the optimal primary treatment end point for late-stage trials has been debated for decades. For pilot and laboratory studies conducted in the context of Phase 1 and early Phase 2 trials, the question of "What outcomes determine safety and efficacy?" has no precise, well-defined answer at this time. Instead, researchers are left to decide which outcomes are most likely to detect safety and efficacy signals in order to increase the probability of moving a medication to late-phase trials. The problems with this approach to medication development are obvious: There is little consistency and consilience between studies in both the measures being administered and variables being assessed. Conversely, the outcome measures for assessing efficacy in Phase 3 trials are well defined. The FDA accepts either complete abstinence or no heavy drinking days as clinical trial end points in order for a medication to be considered for approval. Although both abstinence and low-risk drinking (i.e., consuming alcohol without experiencing a heavy drinking event) after treatment are stable behaviors and associated with positive long-term health outcomes (Kline-Simon et al., 2013, 2014, 2017; Witkiewitz et al., 2017a,b), recent evidence suggests that both end points may be overly strict and may not capture all individuals who respond to a medication.

The Alcohol Clinical Trials Initiative (ACTIVE) workgroup and affiliated researchers have re-examined several of the largest AUD clinical trials to address key issues associated with medication development in this field (Anton et al., 2012). Most relevant to this review, several of such studies have evaluated potential alternate Phase 3 trial end points or

at least extensions of the current FDA-recommended end points Individuals who are able to reduce their alcohol consumption to low-risk levels during treatment do not substantially differ from abstainers in terms of healthcare utilization or medical costs and are able to sustain this reduction in alcohol consumption over several years (Kline-Simon et al., 2014, 2017; Witkiewitz et al., 2017a,b). Furthermore, in a reanalysis of the COMBINE study, individuals who received naltrexone and still reported some heavy drinking days by the end of the study had fewer heavy drinking days than those who received placebo (Falk et al., 2014), consistent with the notion that nonabstainers benefited from naltrexone in COMBINE (Ray et al., 2010b). Similarly, other studies have found that individuals who report limited heavy drinking during treatment have comparable psychosocial functioning to those who report abstinence or nonheavy drinking (Wilson et al., 2016; Witkiewitz et al., 2017a,b). As the individuals described above could be, and often are, categorized as treatment failures by the current FDA-recommended outcomes despite showing beneficial responses to pharmacotherapy, less-stringent drinking outcomes and end points that account for improvements in quality of life should be considered when evaluating medications for AUD.

One measure of alcohol consumption that has received recent attention as a promising alternative trial end point is the World Health Organization (International Guide for Monitoring Alcohol Consumption and Related Harm; WHO, 2000) risk levels of alcohol use (very high risk, high risk, moderate risk, and low risk defined by mean ethanol consumption in grams per day). The European Medicines Agency (EMA), which is Europe's equivalent of the FDA, allows for a 2-level categorical shift in WHO risk levels (e.g., high risk to low risk) as an end point for medication evaluation in addition to complete abstinence, heavy drinking days, and total alcohol consumption (Guideline on the Development of Medicinal Products for the Treatment of Alcohol Dependence; EMA, 2010). For example, WHO risk level reduction was used by the EMA to evaluate, and eventually approve, nalmefene as a treatment for AUD (Selincro Assessment Report; EMA, 2012). Recent studies from the ACTIVE workgroup has supported the validity of WHO risk levels as end point in U.S. studies as well. In another reanalysis of the COMBINE study, reduction in WHO risk level during treatment was predictive of a lower level of alcohol-related consequences and mental health problems at the end of 1 year after treatment (Witkiewitz et al., 2017a). These findings were supported by an examination of National Epidemiologic Survey on Alcohol and Related Conditions cohort data which found that a reduction in WHO risk level was associated with an attenuated risk of alcohol dependence (Hasin et al., 2017). As others have discussed, the addition of WHO drinking risk levels as an end point in U.S. clinical trials could improve medications development not only by assessing another valid and ecologically important marker of medication efficacy but also by facilitating medication development for AUD globally through the creation of common assessment instruments (Johnson, 2017a,b; Litten et al., 2017).

While improving our ability to detect medication effects via implementing additional drinking end points in late-stage (e.g., Phase 3) clinical trials would certainly increase the likelihood of developing new and better pharmacotherapies for AUD, this step would not necessarily affect the issues presently interfering with the success of Phases 1 and 2 trials. Thus, the current review focuses on clinical testing and ways to optimize the process of establishing safety and initial efficacy using both human laboratory and RCT designs. In fact, one of the priorities identified by Litten and colleagues (2012) is to develop and implement efficient screening models for both animal and human laboratory paradigms. In our own work, we have written about the role of human laboratory models of AUD in medications development (Plebani et al., 2012; Ray et al., 2010a; Yardley and Ray, 2017). *Simply put, the central question to be addressed in early efficacy testing is whether a given pharmacotherapy shows sufficient evidence of early efficacy to warrant further testing through randomized controlled trials.* The extent to which the human laboratory paradigms currently available can provide the needed evidence to make these important go/no-go decisions about pharmacotherapies for AUD remains unestablished. Our recent qualitative review of the literature showed little consilience between findings from human laboratory models and findings from clinical trials (Yardley and Ray, 2017). Furthermore, selective reporting biases for laboratory trials which often measure dozens of outcomes likely result in highly inflated estimates of laboratory efficacy (Simmons et al., 2011) and undermines the translational validity of human laboratory methods. Nevertheless, quantitative approaches, such as a meta-analysis, are required to directly address the question of whether human laboratory findings predict clinical trial outcomes; although to our knowledge, valid meta-analytic approaches for testing this translational question have yet to be articulated in the literature. Indeed, this question is particularly complicated and compounded by the fact that the predictive ability of laboratory measures may be specific to a mechanism of action and/or pharmacological target of the compound being tested. Questions about the translational validity of human laboratory methods are amplified for preclinical research where methodological differences are compounded by large biological differences in the test subjects between animal models and clinical subjects (Egli, 2018). While data from different sources and methods using different medications do not lend themselves neatly to integration with current meta-analytic methods, additional efforts to quantify the predictive utility of these models are clearly warranted to test whether laboratory outcomes predict clinical success.

In the field of AUD medications development, we have seen compelling examples of proof-of-concept human laboratory studies that were positive and were followed by a positive result in a RCT. For instance, McKee and colleagues (2009) utilized a laboratory-based self-administration

paradigm which found that varenicline significantly reduced alcohol craving, subjective response to alcohol, and alcohol consumption in heavy drinking smokers (McKee et al., 2009). This initial evidence of efficacy for face-valid laboratory phenotypes was later followed by several small clinical trials (Fucito et al., 2011; Mitchell et al., 2012; Plebani et al., 2013) and a successful RCT which advanced varenicline as an efficacious treatment for alcohol dependence (Litten et al., 2013). However, the field has also seen promising medications based on human laboratory (Ray et al., 2011) and initial clinical trials (Kampman et al., 2007) fail in larger scale confirmatory trials (Litten et al., 2012b). These examples highlight the inconsistent fashion with which results from early efficacy trials, mostly conducted in the human laboratory, relate to clinical outcomes in RCTs, the gold standard for evaluating and ultimately approving novel treatments for AUD.

In recognizing the role of early efficacy studies for medications development, and in particular, of human laboratory studies of AUD pharmacotherapies, the next step is to clearly articulate a set of goals for early efficacy phases of research. To that end, we contend that a useful approach would be to review the literature on the mature field of behavioral therapies development for substance use disorders (SUD). While both fields incorporate a stage model of development, the behavioral therapies approach has been established and utilized for 2 decades (Rounsaville et al., 2001). Therefore, a critical comparison of these models may yield important insights that can, in turn, inform a more efficient and effective course for novel compounds in medications development for AUD.

STAGES OF BEHAVIORAL THERAPIES DEVELOPMENT

The field of behavioral treatment was revolutionized by the advancement of the technology model of research implemented in the National Institute of Mental Health (NIMH) Treatment of Depression Collaborative Research Program (Elkin et al., 1985). This framework provided prescriptive methods for how behavioral techniques should be evaluated, including issues of therapist fidelity and trial monitoring. Building on the technology model of behavioral therapy development and recognizing the large burden on investigators to conduct full-scale efficacy trials, Onken and colleagues (1997) proposed a stage model of behavioral therapies research. Of note, the stage model of behavioral therapies development was preceded by the stage model of pharmacotherapy development and intended to parallel NIDA's medications development program (https://grants.nih.gov/grants/guide/pa-files/PA-94-078.html). In this model, Stage 1 consists of pilot/feasibility testing, which includes research activities such as manual writing, training therapists, and adherence measurement of the novel treatment. Stage 1 is further divided into 2 substages, Stage 1a is described as therapy development and manual writing, while Stage 1b consists of a

pilot trial with the goal of estimating a treatment effect size (Rounsaville et al., 2001). It is not uncommon for Stage 1a to be associated with a small open-label trial to assess the acceptability and feasibility of the novel treatment, prior to enrolling participants into the pilot RCT (Stage 1b). In a more recent description of the stage model of behavioral therapy development, Carroll and Onken (2005) emphasize the notion that Stage 1 research provides the opportunity for creativity and innovation in clinical behavioral science by allowing researchers to develop novel behavioral therapies and/or improve existing treatments. Further, treatment research at this stage has the potential for a high yield with regard to evaluating clinical strategies that have not yet undergone empirical testing. This is also an opportunity to integrate advances in basic neuroscience and basic research on affective, cognitive, and social factors, including brain imaging research into Stage 1a and 1b research (Carroll and Onken, 2005). Stage 2 consists of RCTs to evaluate the efficacy of the pilot-tested treatments and to address mechanistic questions, particularly regarding the mechanisms of action of treatments with established efficacy. Stage 3 consists of studies to evaluate transportability of treatments, including opportunities to make these treatments more usable in community treatment settings (Carroll and Onken, 2005). One notable difference between behavioral treatment development and medications development is that all stages of behavioral treatment development are typically conducted with treatment-seeking samples, likely because fear of adverse events in a vulnerable population is relatively minimal.

The stage model of behavioral therapy research is the "gold standard" in the field of mental health, including addiction, and this has been facilitated by a funding mechanism within the National Institutes of Health dedicated to behavioral therapies development. Specifically, the R34 mechanism supports those research activities proposed for Stage 1 of the behavioral treatment development model. An added advantage of the R34 mechanism is that it allows for 3 years of research support to conduct the proposed research, typically incorporating aspects of Phase 1a and 1b. We are especially interested in the Stage 1 of behavioral therapies research given that it parallels Phase 1 for medications development, namely safety and initial efficacy testing in human subjects (also termed Phases 1a and 1b). To that end, behavioral therapies research has provided concrete guidelines as to what elements are required at each stage of treatment development and how those steps can be integrated into research proposals (Rounsaville et al., 2001). This approach is highly consistent with the technology model of behavioral therapy and has implications for how these therapies are later evaluated for the strength of the evidence of their efficacy (i.e., evidence-based treatments approach; Chambless, 2015).

In summary, a review of the stage model of behavioral treatment development suggests a prescriptive approach for Stage 1 of treatment development, informed largely by the technology model, associated with a discrete mechanism of

NIH support (i.e., R34) and that accommodates a range of clinical research methods and seeks to nurture creativity and innovation. While this model was preceded and influenced by the medications development stage model, the flexibility and clear delineation of research activities at each stage afforded by the behavioral therapies model may in turn be used to refine its predecessor, namely the stage model of medications development. Next, we consider how the behavioral therapies stage model of treatment development could inform an actionable model for bridging the "valley of death" in AUD medications development.

RECOMMENDATIONS FOR AUD MEDICATIONS DEVELOPMENT

The goal of this review was to draw parallels from the well-established stage model of behavioral therapies research for SUD and to discuss its application to research on medications development for AUD. Given the description of each separate model for medications and behavioral therapies development, we take this opportunity to envision how aspects of the behavioral therapies treatment development model may be applied to medications development.

One aspect of the Stage 1 research in behavioral therapies development that is particularly intriguing is the notion of dividing early development functions into Stage 1a and 1b. The medications development model parallels this distinction with targets of safety (1a) and initial efficacy (1b). However, as establishing a safe medication × alcohol interaction is a necessary component of demonstrating that a medication is safe for AUD, laboratory-based alcohol administration studies will often assess safety concurrently with analogues of clinical efficacy such as alcohol craving and subjective responses to alcohol (e.g., Hutchison et al., 2001; Ray et al., 2011, 2017a; Roche et al., 2016). Whether this is a better approach than separating safety and early efficacy testing remains up for debate. The behavioral therapies model would argue that there should be a sequence to the 1a and 1b phases, with one informing the other. In fact, if the acceptability of a behavioral treatment is found to be problematic during the open-label pilot, changes are made to the manual and to the treatment delivery to address those issues (e.g., dosing schedule, compounding in medications development). Separating safety and early efficacy may have an important benefit of allowing the safety trial to generate hypotheses for testing under optimal conditions during Stage 1b activities.

PILOT CLINICAL TRIALS VERSUS LABORATORY STUDIES

Another distinction between the development of behavioral therapies and the development of pharmacotherapies, is the fact that in behavioral therapies development, pilot RCTs are the gold standard for early efficacy testing. Conversely, in medications development, human laboratory models are often the preferred approach to screening novel therapeutics for initial efficacy. As noted previously, there is a rich literature supporting the use of experimental psychopharmacology paradigms in AUD etiology and treatment development (Bujarski and Ray, 2016). An important distinction between behavioral pharmacology studies of AUD medication and clinical trials for AUD is the nature of the sample, with the former being comprised of nontreatment seekers and the latter comprised of treatment-seeking individuals. Our group (Ray et al., 2017b) and others (Rohn et al., 2017) have recently highlighted differences between treatment-seeking and non-treatment-seeking samples on a host of relevant demographic and clinical features. In addition, we have argued that one of the ways to promote consilience between human laboratory models and clinical trials may be to engage a similar population of treatment-seeking individuals (Ray et al., 2017b). To that end, using pilot RCTs with participants who are actively seeking treatment for AUD to screen medications for early efficacy would address concerns associated with sample discrepancies as well as with outcome discrepancies. Concerns about safety with testing medications in affected and thus potentially vulnerable populations represent a central ethical concern in clinical research. To promote efficient medications development, we thus recommend that researchers carefully weigh the 2 opposing considerations of external validity and safety in designing early clinical trials of novel medication.

One important consideration with pilot RCTs concerns the small sample sizes inherent to such studies and particularly whether it is appropriate to use pilot studies to generate effect size estimates for novel treatments. Recent discussion about this issue has suggested that estimating effect sizes from pilot studies is unsupported from a methodological viewpoint (Kraemer et al., 2006; Leon et al., 2011) and that stable estimation of effect size requires hundreds of participants to be enrolled (Schönbrodt and Perugini, 2013). As laboratory studies are typically able to enroll more participants and minimize statistical noise through greater experimental control, the methodological concerns about estimating efficacy are somewhat mitigated. With this caveat, however, it is clear that pilot studies have a lot to offer the literature, including evaluating the feasibility of recruitment, randomization, retention, assessment procedures, and implementation of the novel intervention. Pilot RCTs can also identify modifications needed in the design of a larger hypothesis-testing trial (Leon et al., 2011). In summary, the pilot RCT approach may be better suited at setting the stage for the larger scale efficacy trial (Phase 2) than to establish the initial efficacy of the novel compound. Insofar as grant reviewers and stakeholders have proper expectations for the results of pilot studies, they can play a crucial role in medications development for AUD.

Another important consideration is the fact that investigators may continue to use human laboratory methods to conduct Phase 1b activities in medications development, namely gauging initial efficacy. To the extent that human laboratory models of AUD are suited for medication screening, they

can continue to have a meaningful role in medications development. The use of human laboratory models in medications development for AUD can be bolstered by a set of standardized procedures with common methods and end points that can be implemented reliably across research groups. To that end, NIAAA has established the NIAAA Human Laboratory Program, designed to screen compounds for effectiveness using human laboratory paradigms. The experimental paradigms derived from this important initiative can in turn be validated and disseminated as standards for screening of novel compounds.

A critical direction in using human laboratory models in medications development relates to the identification and refinement of paradigms that have high external validity for predicting clinically relevant (and clinical trial-related) outcomes. In other words, there is a pressing need to develop valid and efficient methods for shepherding novel compounds from initial discovery and safety testing through RCTs and ultimately approval/dissemination. While traditional human laboratory methods have been proposed to serve these aims, at present, no human laboratory method has demonstrated translational validity. The degree to which human laboratory outcomes predict AUD clinical trial outcomes remains hotly debated yet mostly untested (Litten et al., 2016; Ray et al., 2016; Yardley and Ray, 2017). In short, human laboratory studies have been measuring the same core of efficacy outcomes (e.g., how a medication affects subjective response to alcohol, alcohol cue-induced craving, and alcohol self-administration), albeit with improving measurement precision, for decades with minimal advancement or variation. The outcomes from Phase 1 and 2 medication testing should provide confidence in the success of Phase 3 trials FDA-recommended clinical trial end points. Despite the lengthy list of experimental psychopharmacology paradigms used in AUD research (Bujarski and Ray, 2016), as a field, we have no evidence that a medication's effect on any of these measures is predictive of that medication's success in Phase 3 trials. Let us examine subjective response to alcohol as an example. The subjective stimulating/rewarding and sedative/aversive effects of alcohol are predictive of AUD development and drinking behavior (e.g., King et al., 2011). Thus, as a logical extension, it has been widely and implicitly accepted without explicit evidence that a medication's ability to block the rewarding effects or potentiate the aversive effects of alcohol in the laboratory transitively confers the ability to reduce hazardous drinking in the real world. Yet, in medication development for other addictive substances, the utility of subjective response to acute drug administration as an outcome measure appears to be limited (Comer et al., 2008; Haney, 2009). For example, the dose at which a medication effectively reduces positive subjective responses to cocaine is not necessarily sufficient to reduce self-administration in the laboratory or drug use in a clinical trial setting (Comer et al., 2008; Haney and Spealman, 2008). Given the obstacles to medication development discussed in this review, we believe it is

imperative that future AUD research employs combined laboratory longitudinal approaches under the umbrella of late Phase 2 or Phase 3 trials to establish the predictive utility of pharmacological manipulation of laboratory-based methods as a robust and reliable indicator of medication efficacy in clinical settings. Additionally, we should seek to validate new human laboratory paradigms that increase confidence in future success in Phase 3 trials. A model seeking to bolster external validity of medication screening in the laboratory has been tested in the smoking literature with early success (Perkins and Lerman, 2014; Perkins et al., 2006, 2008, 2010). This model consists of a "practice quit attempt" in which smokers are asked to abstain for 1 week and to come to the laboratory daily for bioverification of nicotine abstinence and would seemingly translate quite well to AUD samples (Perkins and Lerman, 2014). In brief, the more that the human laboratory paradigm can approximate drinking outcomes, the more likely it is to provide valuable information regarding the initial efficacy of a novel compound.

SIMULATION COMPARING PILOT TRIALS TO LABORATORY STUDIES

To further explore the relative efficiency of pilot trials versus laboratory studies in screening medications for AUD, we conducted a Monte Carlo simulation study. For this simulation study, we tested the following parameters: (i) average medication effect size (Cohen's d = 0.2, 0.5, and 0.8 representing small, medium, and large effects), pilot study sample size (N_{pilot} range 6 to 36), the multiplicative increase in sample size associated with typically less expensive and quicker laboratory studies (Lab Multiple = 1, 2, and 4), and lastly the correlation between clinical and laboratory effect sizes ($\rho_{Lab\text{-}Clinic}$ = 0.3, 0.6, 0.9, representing relatively poor, moderate, and strong correlation between clinical and laboratory effects). For each of these parameter combinations, we simulated 10,000 clinical effect sizes, half from a null distribution $N(0, 0.2)$ and half from the specified mean effect size $N(d, 0.2)$. Laboratory effect sizes were then simulated based on population correlation $\rho_{Lab\text{-}Clinic}$ with equivalent distributions to the clinical effect sizes. We then calculated the probability of a significant and positive trial (i.e., statistical power with α = 0.025, 1-tailed) for each of the simulated effect sizes in both a pilot study of sample size N_{pilot} and a laboratory study of sample size $N_{pilot} \times$ Lab Multiple. By summing these powers, we calculated the expected number of positive trials for medications with true and false effects (i.e., drawn from distribution $N(d, 0.2)$ or $N(0, 0.2)$, respectively). We then calculated the sensitivity and positive predictive value for each parameter combination. All Monte Carlo simulations were conducted in R by SB. Full R code and all figures are available at www.github.com/sbujarski.

As early efficacy trials are primarily interested in testing whether a treatment shows sufficient evidence of efficacy to
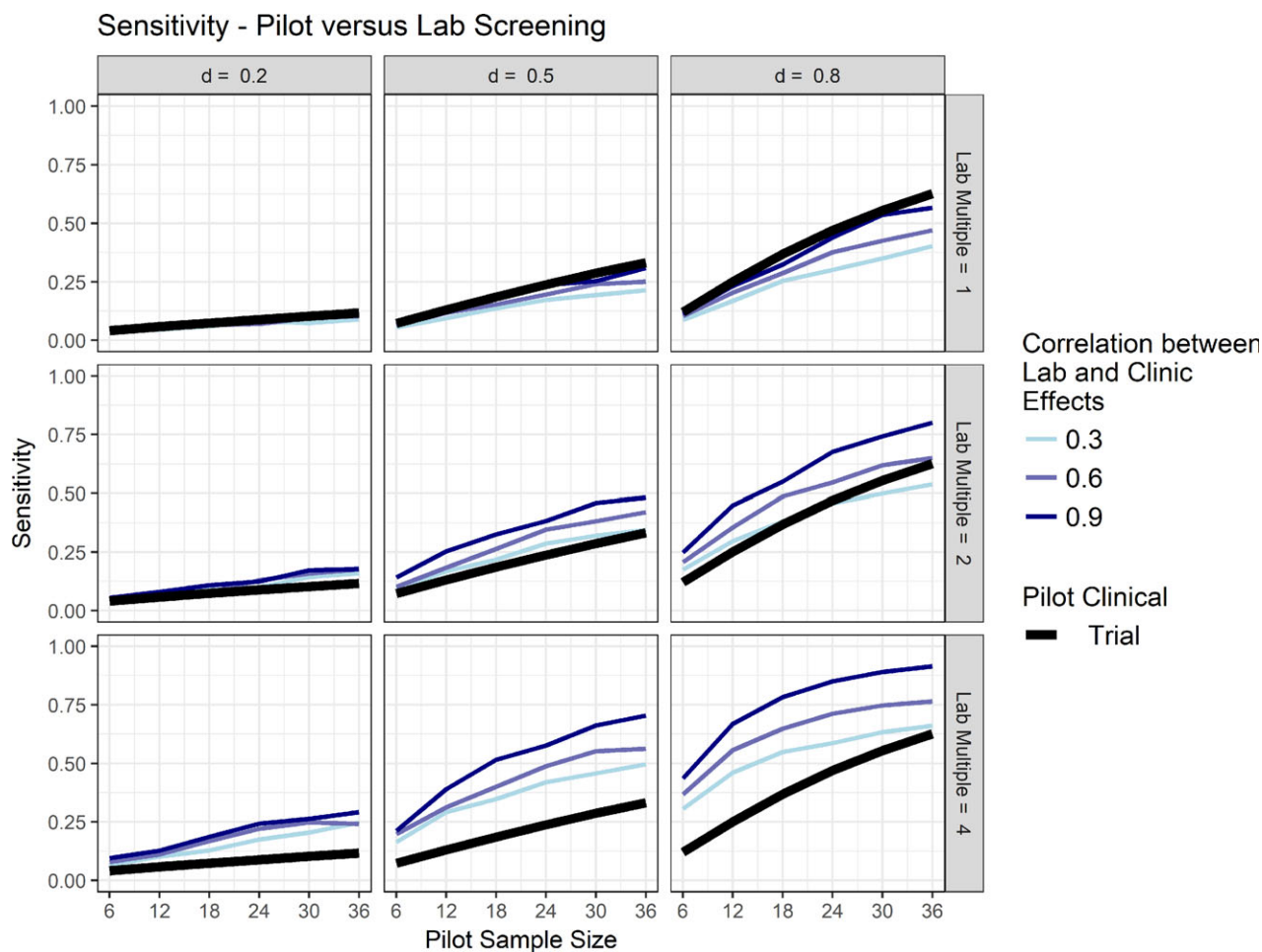
Sensitivity - Pilot versus Lab Screening

Fig. 1. Sensitivity results for pilot versus laboratory screening of novel medications. These sensitivity analyses are the result of a Monte Carlo simulation study (n = 10,000 per specification) examining the effect of pilot study sample size, mean medication effect size (Cohen's d), multiplicative increase in laboratory versus pilot clinical trial (Lab Multiple), and correlation between laboratory and clinic effect sizes. Sensitivity was defined as the expected number of positive screens as a proportion of the total number of true positive medications.

warrant a full-scale Phase 3 RCT, we focused primarily on sensitivity or the probability that a medication with a true effect would screen positive. As expected, results showed that sensitivity increased with greater sample size, greater average effect size, and greater correlations between laboratory and clinic effect sizes (see Fig. 1).

These simulations suggest that a pilot study is superior in terms of sensitivity only when a laboratory study confers no advantage in terms of sample size (i.e., Lab Multiple = 1). However, when a laboratory study has twice the sample size of a pilot study, even a mediocre laboratory paradigm (i.e., $\rho_{LabClinic}$ = 0.6) exhibited greater sensitivity than a pilot trial across the range of sample sizes and medication effect sizes tested. For laboratory studies enrolling 4 times the sample size, sensitivity was superior to pilot studies even with a poor laboratory analogue ($\rho_{LabClinic}$ = 0.3). However, while laboratory studies had generally greater sensitivity, pilot studies were generally superior to laboratory studies in terms of positive predictive value or the probability that a positive trial is a true positive (Fig. 2). As the probability of a statistically significant pilot trial is so low, a significant result is more

predictive of a true effect as compared to more highly powered laboratory study which measures a more distal outcome. It is worth noting, however, that the effect sizes calculated in small, yet statistically significant, studies are likely to be dramatically inflated and thus should not be used to determine sample sizes for large-scale RCTs (Gelman and Carlin, 2014).

In sum, this Monte Carlo simulation found that a laboratory study with a paradigm well calibrated to capture meaningful clinical effects and do so with a larger sample size is better able to detect true positive medication effects. However, as the concordance between laboratory and clinical effects is not perfect, positive signals are less likely to be predictive of a true positive medication effect than the effects obtained in pilot trials. Furthermore, these simulations reinforce the conclusions of Leon and colleagues (2011) and Kraemer and colleagues (2006) that small pilot trials are ill-equipped to estimate effect sizes or even engage in hypothesis testing of treatment efficacy. Laboratory studies that enroll more subjects may be better positioned to engage in hypothesis testing, but this assumes that the laboratory outcomes
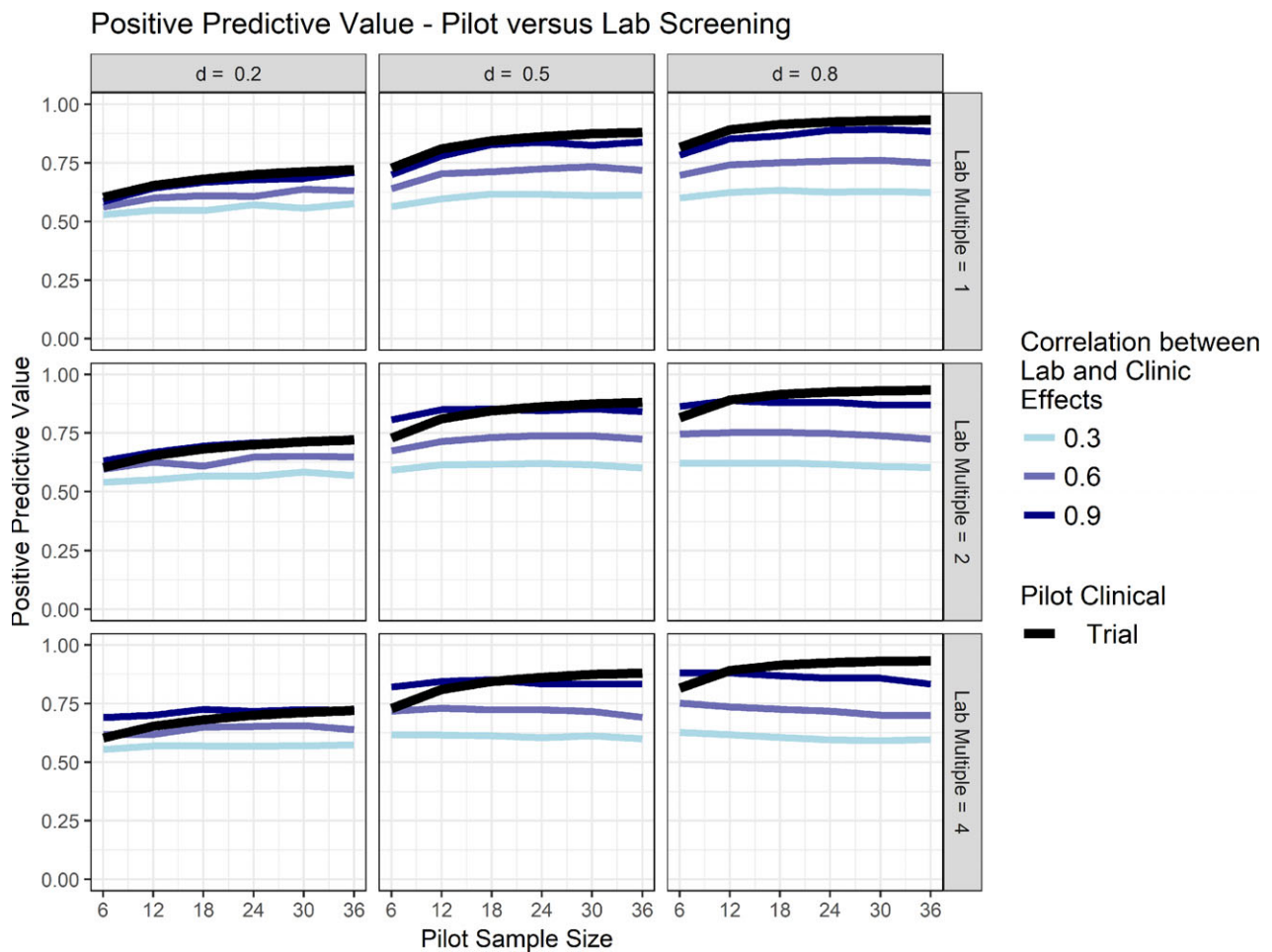
Fig. 2. Positive predictive value results for pilot versus laboratory screening of novel medications. These positive predictive values are the result of a Monte Carlo simulation study ($n = 10,000$ per specification) examining the effect of pilot study sample size, mean medication effect size (Cohen's d), multiplicative increase in laboratory versus pilot clinical trial (Lab Multiple), and correlation between laboratory and clinic effect sizes. Positive predictive value was defined as the proportion of positive screens that represent true positives.

correlate with clinical efficacy, an assertion that is currently unproven in the field. Last, it should be noted that results of this simulation, while intended to inform medication testing for AUD, would be generalizable and applicable to medication development for other psychiatric and medical conditions that use the FDA Clinical Research Phases.

STRUCTURAL SUPPORT FOR EARLY-PHASE TESTING

Upon reading the series of stages in behavioral treatment development, pharmacology researchers may quickly raise issues of time and feasibility. Phase 1 activities are very involved and time-consuming in the context of medications development, although still less in comparison than Phases 2 and 3 testing. The recognition of the burden on investigators to implement clinical trials within the framework of the technology model led NIMH to recognize manual development and initial open-label testing as important activities that require time and support. To that end, the 3-year scope of the R34 mechanism may provide a unique set of benefits to

the behavioral therapies approach, which are currently unavailable to medications development. This is particularly relevant with increasing recognition of the regulatory burden associated with medications development, including tasks such as obtaining and managing an IND application to the FDA, implementing good clinical practices across all study procedures, and establishing and maintaining regular and productive communications with various stakeholders such as pharmaceutical companies, Institutional Review Boards, university entities, clinical research centers, and research participants. Taken together, these regulatory and research activities argue for increased support for Phase I activities in medications development as a way to overcome the "valley of death" and to more effectively move promising compounds from preclinical to clinical studies.

CONCLUSIONS

Medications development for AUD is a high research priority area with great potential to have a meaningful public health impact. As with other fields of science, developing

novel therapeutics for AUD is not without a host of obstacles. A notable obstacle in medications development for AUD has been the transition from preclinical testing to human testing, which has been termed the "valley of death," given that very few promising therapeutics from preclinical studies are actually tested in humans. To overcome this notable obstacle, this review focuses on the first steps in human testing, particularly studies of safety and initial efficacy of a novel compound. Regulatory issues aside, we discuss ways in which initial efficacy testing can take place and draw upon the literature on behavioral therapies development, which more clearly prescribes a pilot RCT as an intermediate step between the development of a novel behavioral treatment and an efficacy trial (i.e., RCT). In the behavioral pharmacology tradition, however, human laboratory end points (e.g., alcohol craving and subjective response to alcohol) are often used as indicators of early efficacy. This review contrasts the human laboratory and pilot RCT approaches both in theoretical terms and using a Monte Carlo data simulation approach. The pilot RCT approach is advantageous over the human laboratory approach in terms of ecological validity as well as for determining treatment tolerability and acceptability, whereas the human laboratory approach appears to be more cost-effective and time-efficient. A major limitation of the pilot RCT approach is that outcomes from small trials should not be used to estimate effect sizes, whereas a major limitation of the human laboratory approach is that their outcomes may not be robustly associated with the clinical outcomes we are most interested in. Results of our Monte Carlo simulation found that a laboratory study with a paradigm well calibrated to capture meaningful clinical effects is better able to detect true positive medication effects than pilot RCTs. Together, these results argue for the careful "calibration" of human laboratory paradigms as a critical step in overcoming the "valley of death" in medications development for AUD and ultimately promoting the translation of novel compounds to clinical populations. However, structural and funding changes are necessary to address barriers to early-stage medication development. In particular, we have highlighted the importance of the R34 mechanism in behavioral treatment development research in supporting scientists in the early stages of human subjects testing which for medications development carries a high regulatory burden including IND applications, developing practice guidelines, and coordination with various stakeholders. By providing grant support for these laborious yet necessary steps in medications development, the NIH could encourage scientists to shepherd a medication across the "valley of death."

ACKNOWLEDGMENTS

CONFLICT OF INTEREST

REFERENCES

Anton RF, Litten RZ, Falk DE, Palumbo JM, Bartus RT, Robinson RL, Kranzler HR, Kosten TR, Meyer RE, O'Brien CP, Mann K, Meulien D (2012) The Alcohol Clinical Trials Initiative (ACTIVE): purpose and goals for assessing important and salient issues for medications development in alcohol use disorders. Neuropsychopharmacology 37:402–411.

Blodgett JC, Del Re AC, Maisel NC, Finney JW (2014) A meta-analysis of topiramate's effects for individuals with alcohol use disorders. Alcohol Clin Exp Res 38:1481–1488.

Bujarski S, Ray LA (2016) Experimental psychopathology paradigms for alcohol use disorders: applications for translational research. Behav Res Ther 86:11–22.

Carroll KM, Onken LS (2005) Behavioral therapies for drug abuse. Am J Psychiatry 162:1452–1460.

Chambless DL (2015) Bringing identification of empirically supported treatments into the 21st century. Clin Psychol Sci Pr 22:339–342.

Comer SD, Ashworth JB, Foltin RW, Johanson CE, Zacny JP, Walsh SL (2008) The role of human drug self-administration procedures in the development of medications. Drug Alcohol Depend 96:1–15.

Donoghue K, Elzerbi C, Saunders R, Whittington C, Pilling S, Drummond C (2015) The efficacy of acamprosate and naltrexone in the treatment of alcohol dependence, Europe versus the rest of the world: a meta-analysis. Addiction 110:920–930.

Egli M (2005) Can experimental paradigms and animal models be used to discover clinically effective medications for alcoholism? Addict Biol 10:309–319.

Egli M (2018) Advancing pharmacotherapy development from preclinical animal studies. Handb Exp Pharmacol https://doi.org/10.1007/164_2017_85 [Epub ahead of print].

Elkin I, Parloff MB, Hadley SW, Autry JH (1985) NIMH treatment of depression collaborative research program. Background and research plan. Arch Gen Psychiatry 42:305–316.

European Medicines Agency (2010) Guideline on the development of medicinal products for the treatment of alcohol dependence. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/03/WC500074898.pdf. Accessed July 18, 2018.

European Medicines Agency (2012) Assessment Report: Selincro. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/002583/WC500140326.pdf. Accessed July 18, 2018.

Falk DE, Litten RZ, Anton RF, Kranzler HR, Johnson BA, Active W (2014) Cumulative proportion of responders analysis (CPRA) as a tool to assess treatment outcome in alcohol clinical trials. J Stud Alcohol Drugs 75:335–346.

Fucito LM, Toll BA, Wu R, Romano DM, Tek E, O'Malley SS (2011) A preliminary investigation of varenicline for heavy drinking smokers. Psychopharmacology 215:655–663.

Fuller RK, Hiller-Sturmhofel S (1999) Alcoholism treatment in the United States. An overview. Alcohol Res Health 23:69–77.

Gelman A, Carlin J (2014) Beyond power calculations: assessing type S (Sign) and type M (Magnitude) errors. Perspect Psychol Sci 9:641–651.

Haney M (2009) Self-administration of cocaine, cannabis and heroin in the human laboratory: benefits and pitfalls. Addict Biol 14:9–21.

Haney M, Spealman R (2008) Controversies in translational research: drug self-administration. Psychopharmacology 199:403–419.

Hasin DS, Wall M, Witkiewitz K, Kranzler HR, Falk D, Litten R, Mann K, O'Malley SS, Scodes J, Robinson RL, Anton R; Initiative ACT (2017) Change in non-abstinent WHO drinking risk levels and alcohol

dependence: a 3 year follow-up study in the US general population. Lancet Psychiat 4:469–476.

Hutchison KE, Swift R, Rohsenow DJ, Monti PM, Davidson D, Almeida A (2001) Olanzapine reduces urge to drink after drinking cues and a priming dose of alcohol. Psychopharmacology 155:27–34.

Johnson BA (2017a) FDA and EMA need homology on alcohol outcome measures—*Semper: Simplicitas est purius modum*. Alcohol Clin Exp Res 41:1383–1384.

Johnson BA (2017b) Toward rational, evidence-based, and clinically relevant measures to determine improvement following treatment for alcohol use disorder. Alcohol Clin Exp Res 41:703–707.

Jonas DE, Amick HR, Feltner C, Bobashev G, Thomas K, Wines R, Kim MM, Shanahan E, Gass CE, Rowe CJ, Garbutt JC (2014) Pharmacotherapy for adults with alcohol use disorders in outpatient settings: a systematic review and meta-analysis. JAMA 311:1889–1900.

Kampman KM, Pettinati HM, Lynch KG, Whittingham T, Macfadden W, Dackis C, Tirado C, Oslin DW, Sparkman T, O'Brien CP (2007) A double-blind, placebo-controlled pilot trial of quetiapine for the treatment of Type A and Type B alcoholism. J Clin Psychopharmacol 27:344–351.

King AC, de Wit H, McNamara PJ, Cao D (2011) Rewarding, stimulant, and sedative alcohol responses and relationship to future binge drinking. Arch Gen Psychiatry 68:389–399.

Kline-Simon AH, Falk DE, Litten RZ, Mertens JR, Fertig J, Ryan M, Weisner CM (2013) Posttreatment low-risk drinking as a predictor of future drinking and problem outcomes among individuals with alcohol use disorders. Alcohol Clin Exp Res 37(Suppl 1):E373–E380.

Kline-Simon AH, Litten RZ, Weisner CM, Falk DE (2017) Posttreatment low-risk drinking as a predictor of future drinking and problem outcomes among individuals with alcohol use disorders: a 9-year follow-up. Alcohol Clin Exp Res 41:653–658.

Kline-Simon AH, Weisner CM, Parthasarathy S, Falk DE, Litten RZ, Mertens JR (2014) Five-year healthcare utilization and costs among lower-risk drinkers following alcohol treatment. Alcohol Clin Exp Res 38:579–586.

Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA (2006) Caution regarding the use of pilot studies to guide power calculations for study proposals. Arch Gen Psychiat 63:484–489.

Leon AC, Davis LL, Kraemer HC (2011) The role and interpretation of pilot studies in clinical research. J Psychiatr Res 45:626–629.

Litten RZ, Egli M, Heilig M, Cui C, Fertig JB, Ryan ML, Falk DE, Moss H, Huebner R, Noronha A (2012a) Medications development to treat alcohol dependence: a vision for the next decade. Addict Biol 17:513–527.

Litten RZ, Falk DE, O'Malley SS, Witkiewitz KA, Mann KF, Anton RF (2017) Letter to editor in response to Johnson's commentary (2017) on the Witkiewitz and colleagues (2017) article. Alcohol Clin Exp Res 41:1381–1382.

Litten RZ, Falk D, Ryan M, Fertig J (2014a) Research opportunities for medications to treat alcohol dependence: addressing stakeholders' needs. Alcohol Clin Exp Res 38:27–32.

Litten RZ, Falk DE, Ryan ML, Fertig JB (2016) Discovery, development, and adoption of medications to treat alcohol use disorder: goals for the phases of medications development. Alcohol Clin Exp Res 40:1368–1379.

Litten RZ, Fertig JB, Falk DE, Ryan ML, Mattson ME, Collins JF, Murtaugh C, Ciraulo D, Green AI, Johnson B, Pettinati H, Swift R, Afshar M, Brunette MF, Tiouririne NA, Kampman K, Stout R (2012b) A double-blind, placebo-controlled trial to assess the efficacy of quetiapine fumarate XR in very heavy-drinking alcohol-dependent patients. Alcohol Clin Exp Res 36:406–416.

Litten RZ, Ryan M, Falk D, Fertig J (2014b) Alcohol medications development: advantages and caveats of government/academia collaborating with the pharmaceutical industry. Alcohol Clin Exp Res 38:1196–1199.

Litten RZ, Ryan ML, Falk DE, Reilly M, Fertig JB, Koob GF (2015) Heterogeneity of alcohol use disorder: understanding mechanisms to advance personalized treatment. Alcohol Clin Exp Res 39:579–584.

Litten RZ, Ryan ML, Fertig JB, Falk DE, Johnson B, Dunn KE, Green AI, Pettinati HM, Ciraulo DA, Sarid-Segal O, Kampman K, Brunette MF, Strain EC, Tiouririne NA, Ransom J, Scott C, Stout R (2013) A double-blind, placebo-controlled trial assessing the efficacy of varenicline tartrate for alcohol dependence. J Addict Med 7:277–286.

McKee SA, Harrison EL, O'Malley SS, Krishnan-Sarin S, Shi J, Tetrault JM, Picciotto MR, Petrakis IL, Estevez N, Balchunas E (2009) Varenicline reduces alcohol self-administration in heavy-drinking smokers. Biol Psychiatry 66:185–190.

Mitchell JM, Teague CH, Kayser AS, Bartlett SE, Fields HL (2012) Varenicline decreases alcohol consumption in heavy-drinking smokers. Psychopharmacology 223:299–306.

Onken LS, Blaine JD, Battjes R (1997) Behavioral therapy research: a conceptualization of a process, in *Innovative Approaches from Difficult to Treat Populations, Innovative Approaches from Difficult to Treat Populations* (Henngler SW, Amentos R eds), pp 477–485. American Psychiatric Press, Washington, DC.

Perkins KA, Lerman C (2014) An efficient early phase 2 procedure to screen medications for efficacy in smoking cessation. Psychopharmacology 231:1–11.

Perkins KA, Lerman C, Fonte CA, Mercincavage M, Stitzer ML, Chengappa KN, Jain A (2010) Cross-validation of a new procedure for early screening of smoking cessation medications in humans. Clin Pharmacol Ther 88:109–114.

Perkins KA, Lerman C, Stitzer M, Fonte CA, Briski JL, Scott JA, Chengappa KN (2008) Development of procedures for early screening of smoking cessation medications in humans. Clin Pharmacol Ther 84:216–221.

Perkins KA, Stitzer M, Lerman C (2006) Medication screening for smoking cessation: a proposal for new methodologies. Psychopharmacology 184:628–636.

Plebani JG, Lynch KG, Rennert L, Pettinati HM, O'Brien CP, Kampman KM (2013) Results from a pilot clinical trial of varenicline for the treatment of alcohol dependence. Drug Alcohol Depend 133:754–758.

Plebani JG, Ray LA, Morean ME, Corbin WR, MacKillop J, Amlung M, King AC (2012) Human laboratory paradigms in alcohol research. Alcohol Clin Exp Res 36:972–983.

Ray LA, Bujarski S, Roche DJ (2016) Subjective response to alcohol as a research domain criterion. Alcohol Clin Exp Res 40:6–17.

Ray LA, Bujarski S, Shoptaw S, Roche DJ, Heinzerling K, Miotto K (2017a) Development of the neuroimmune modulator ibudilast for the treatment of alcoholism: a randomized, placebo-controlled, human laboratory trial. Neuropsychopharmacology 42:1776–1788.

Ray LA, Bujarski S, Yardley MM, Roche DJO, Hartwell EE (2017b) Differences between treatment-seeking and non-treatment-seeking participants in medication studies for alcoholism: do they matter? Am J Drug Alcohol Abuse 43:703–710.

Ray LA, Chin PF, Heydari A, Miotto K (2011) A human laboratory study of the effects of quetiapine on subjective intoxication and alcohol craving. Psychopharmacology 217:341–351.

Ray LA, Hutchison KE, Tartter M (2010a) Application of human laboratory models to pharmacotherapy development for alcohol dependence. Curr Pharm Des 16:2149–2158.

Ray LA, Krull JL, Leggio L (2010b) The effects of naltrexone among alcohol non-abstainers: results from the COMBINE study. Front Psychiatry 1:26.

Ray LA, Roche DJ, Heinzerling K, Shoptaw S (2014) Opportunities for the development of neuroimmune therapies in addiction. Int Rev Neurobiol 118:381–401.

Roche DJ, Yardley MM, Lunny KF, Louie SG, Davies DL, Miotto K, Ray LA (2016) A pilot study of the safety and initial efficacy of ivermectin for the treatment of alcohol use disorder. Alcohol Clin Exp Res 40:1312–1320.

Rohn MC, Lee MR, Kleuter SB, Schwandt ML, Falk DE, Leggio L (2017) Differences between treatment-seeking and nontreatment-seeking alcohol-dependent research participants: an exploratory analysis. Alcohol Clin Exp Res 41:414–420.

Rösner S, Hackl-Herrwerth A, Leucht S, Lehert P, Vecchi S, Soyka M (2010) Acamprosate for alcohol dependence. Cochrane Database Syst Rev 9:CD004332.

Rounsaville BJ, Carroll KM, Onken LS (2001) A stage model of behavioral therapies research: getting started and moving on from stage I. Clin Psychol Sci Pr 8:133–142.

Schönbrodt FD, Perugini M (2013) At what sample size do correlations stabilize? J Res Pers 47:609–612.

Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 22:1359–1366.

Wilson AD, Bravo AJ, Pearson MR, Witkiewitz K (2016) Finding success in failure: using latent profile analysis to examine heterogeneity in psychosocial functioning among heavy drinkers following treatment. Addiction 111:2145–2154.

Witkiewitz K, Hallgren KA, Kranzler HR, Mann KF, Hasin DS, Falk DE, Litten RZ, O'Malley SS, Anton RF (2017a) Clinical validation of reduced alcohol consumption after treatment for alcohol dependence using the world health organization risk drinking levels. Alcohol Clin Exp Res 41:179–186.

Witkiewitz K, Wilson AD, Pearson MR, Hallgren KA, Falk DE, Litten RZ, Kranzler HR, Mann KF, Hasin DS, O'Malley SS, Anton RF (2017b) Temporal stability of heavy drinking days and drinking reductions among heavy drinkers in the COMBINE study. Alcohol Clin Exp Res 41:1054–1062.

World Health Organization (2000) International guide for monitoring alcohol consumption and related harm. WHO, Geneva, Switzerland.

Yardley MM, Ray LA (2017) Medications development for the treatment of alcohol use disorder: insights into the predictive value of animal and human laboratory models. Addict Biol 22:581–615.