

Analysis of Extreme Drinking in Patients with Alcohol Dependence Using Pareto Regression

Sourish Das¹, Ofer Harel², Dipak Dey², Jonathan Covault³, Henry Kranzler³

¹ SAMSI and Duke University.

² Department of Statistics, University of Connecticut.

³ Department of Psychiatry, University of Connecticut Health Center.

October 23, 2008

Abstract

We developed a novel Pareto regression model with unknown shape parameter to analyze extreme drinking in patients with Alcohol Dependence (AD). We used a generalized linear models (GLM) framework and a log-link between the shape parameter of the random and systematic components and a Monte Carlo based Bayesian method to implement the analysis. We examined two issues of importance in the study of AD: First, we tested whether a single nucleotide polymorphism within *GABRA2* gene, which encodes a subunit of the *GABA_A* receptor and has been associated to AD, influenced extreme alcohol intake and second, the efficacy of three psychotherapies for alcoholism in treating extreme drinking behavior. European-American participants (n = 812, 73.4% male) from Project MATCH, a multi-center randomized clinical trial of the psychotherapeutic treatment of alcoholism, provided DNA samples for this study. Following 3-month treatment period, during which patients received one of the three

psychotherapy treatment, participants were followed up at 3 month intervals. We found that women with the high-risk *GABRA2* allele had a significantly higher probability of extreme drinking behavior than women with no high-risk allele. High-risk women also responded to therapy better than those with two low-risk alleles. We found that women who received *cognitive behavioral therapy* had better outcomes than those those receiving either of the other two therapies. Among men, there was no significant effect of *GABRA2* genotype on extreme drinking behavior. However, *motivational enhancement therapy* was the best treatment for the extreme drinking behavior.

Key words: Alcohol Dependence, Bayesian Modeling, Extreme Behavior, *GABRA2*, Jeffreys' prior, Pareto Regression.

1 Introduction

Alcohol dependence (AD), a disorder associated with significant adverse medical and psychosocial consequences, has been shown to be 50 – 60% heritable (Kendler, 2001). Bauer *et al.* (2007) demonstrated an association between a single nucleotide polymorphism (SNP) within *GABRA2*, the gene encoding the α -2 subunit of the *GABA_A* receptor, and the probability of daily drinking and heavy drinking behavior. In this paper, we examined whether a SNP within *GABRA2* affects extreme alcohol intake and the efficacy of three psychotherapy treatments for alcoholism in treating extreme drinking behavior. An association between *GABRA2* and extreme drinking behavior would help to identify the cause of this high-risk behavior.

We develop a Pareto regression model with unknown shape parameter to analyze extreme drinking behavior of patients diagnosed with AD. We used a *generalized linear models* (GLM) framework and a log-link between the shape parameter of the random and systematic

components. We present a Monte Carlo based Bayesian method to implement the analysis. The format of the paper is as follows. In section 2, we present the study description and data structure. In section 3, we present the Pareto regression model for extreme drinking and a simple Monte Carlo based Bayesian method to analyze the data with extreme response. In section 4, we analyze and present the data from patients in treatment for AD. Section 5 concludes the paper with a brief discussion of these findings.

2 Study Description and Data Structure.

The present study examined a SNP marker in the AD-associated haplotype block in *GABRA2*, the gene encoding the α -2 subunit of the *GABA_A* receptor (Covault et al. 2004) using DNA samples from AD patients who participated in Project MATCH (Matching Alcoholism Treatment to Client Heterogeneity), a psychotherapy study conducted at 10 sites around the United States (Project MATCH research group, 1997 a,b). The project included two parallel, but independent, clinical arms in which patients were recruited from outpatient settings ($n = 952$; 72% male) or aftercare settings following inpatient or day hospital treatment ($n = 774$; 80% male). In each of the two arms patients were randomly assigned to one of the three manual-guided, individually delivered treatments: Motivational Enhancement Therapy (MET), Cognitive Behavioral Therapy (CBT), or Twelve-Step Facilitation (TSF). In this study, we focus on the subsample of 812 European-American patients who provided a DNA sample for analysis. The study length included a 3-month treatment period and a 12-month post treatment period.

The data structure is presented in Table 1. Each row represents data for a single patient. For example, y_{11} is the number of standard drinks ingested by the first patient in the data set that exceeded a threshold (say u) of heavy drinking on day t_{11} of the study period. Similarly, y_{ij} is the number of standard drinks consumed by the i^{th} patient that exceeded a threshold u

of heavy drinking on day t_{ij} of the study period. The last column of the table \mathbf{x}_i is a vector of all of the covariate information, including *age*, *GABRA2* (i.e., genotype), CBT, MET, TSF (i.e., three psychotherapy conditions), and gender.

3 Pareto Regression Model for Extreme Drinking.

Heavy drinking is defined as 5 or more standard drinks per day for men and 4 or more standard drinks per day for women. A standard drink is 0.5 oz of ethanol (See NIAAA 2005). Following this definition of heavy drinking, we modeled drinking behavior as exceeding these thresholds. In practice, tail data for the exceedances over high thresholds are often modeled by fitting a generalized Pareto distribution. There is a rich corresponding literature on the statistics of extreme value. The first publication on the topic was by Gnedenko (1943), which was followed by Pickand (1975). More recent efforts include those by Davison and Smith (1990), Coles and Tawn (1996), Behrens, Lopes and Gamerman (2004), Tancredi, Anderson and O'Hagan (2006) and Castellanos and Cabras (2007), who developed statistical models for univariate extreme value theory for exceedances over thresholds. Here we present a method to examine the effects of covariate information on extreme values of exceedances over thresholds such as extreme drinking. Recently, Tsionas (2003) developed a hierarchical Pareto regression model using a generalized Pareto distribution for a macroeconomic data set using Markov Chain Monte Carlo (MCMC) techniques (see Robert et al. 2005). Here we present a Monte Carlo based Bayesian method to analyze a Pareto regression model for extreme drinking behavior.

A hierarchical Pareto regression model can be viewed as a class of general linear model. In general, GLMs have three components: viz. (i) a random component, (ii) a systematic component and (iii) a link function. We define the Pareto regression model as follows:

Random Component Let Y be the number of standard drinks (exceeding the thresholds of heavy drinking) consumed by a patient on a particular day with a Pareto distribution, say $Pareto(u, \alpha_{ij})$ with cumulative density function as

$$G(y_{ij}) = P(Y > y_{ij} | Y \geq u) = (\tau y_{ij})^{-\alpha_{ij}},$$

where $j = 1, \dots, n_i$, $i = 1, 2, \dots, k$, and $\tau = \frac{1}{u}$ is a known positive constant and $\alpha_{ij} > 0$ is an unknown shape parameter. Following NIAAA(2005) we chose $u = 5$ for men and $u = 4$ for women.

Systematic Component We incorporate the covariate information through a systematic component of GLM,

$$\eta_{ij} = x'_{ij}\beta, \quad j = 1, 2, \dots, n_i \text{ and } i = 1, 2, \dots, k.$$

Link function We use the log-link function as

$$\eta_{ij} = \log(\alpha_{ij}), \quad i = 1, 2, \dots, k \text{ and } j = 1, 2, \dots, n_i.$$

Combining all three components of GLM, the Pareto regression model reduces to

$$P(Y > y_{ij} | Y \geq u) = (\tau y_{ij})^{-\exp\{\beta_0 + \beta_1 x_{1ij} + \dots + \beta_p x_{pij}\}}.$$

Note that we measure the effect of extreme drinking that exceeds the heavy drinking threshold as a continuous measure (as 0.5 oz of ethanol is defined as a standard drink and the number of standard drinks can take a value up to two decimal places; see NIAAA 2005). This contrasts with the binary measures used by Bauer et al. (2007).

3.1 Pareto Regression for Extreme Drinking of AD patients

Despite a substantial sample size, the number of subjects and the frequency of minor allele (i.e., less common) were not sufficient to support an analysis wherein genotype could be

stratified by 3 levels, given the need to stratify by 3 levels of treatment assignment and 2 levels of sex. Instead, genotype was stratified by 2 levels: homozygotes for the low AD risk A-allele (i.e., A/A, which comprises the low-risk genotype group) versus carriers of the AD-associated G-allele (i.e., A/G or G/G, which comprise the high-risk genotype group). This strategy is consistent with the grouping scheme employed in prior studies (Covault et al., 2004; Lappalainen et al., 2005) which showed an association of the G-allele with risk of AD, as well as the response to the acute effects of alcohol among moderate drinkers (Pierucci-Lagha et al. 2005).

The Pareto regression model for the extreme drinking of AD patients is

$$\begin{aligned}
\log(\alpha_{ij}) = & \beta_1 + \beta_2 \log(t_{ij}) + \beta_3 \log(\text{Age}_{ij}) \\
& + \beta_4 I(\text{High risk genotype of } GABRA2 \text{ gene}) \\
& + \beta_5 \log(t_{ij}) \times I(\text{High risk genotype of } GABRA2 \text{ gene}) \\
& + \beta_6 I(\text{patient treated under CBT}) \\
& + \beta_7 I(\text{patient treated under MET}),
\end{aligned}$$

where $I(A) = 1(0)$, if A is true (otherwise), is an indicator variable. Since the definition of heavy drinking differs by sex, we fit a separate model for each sex.

3.2 Bayesian Analysis

In analyzing complex models using Bayesian techniques, we often generate MCMC posterior samples of the target parameters (here regression parameters) β , using Gibbs or Metropolis-Hastings algorithms (see Robert et al. 2005). We present a Monte Carlo technique, (Das 2008) to implement the Bayesian analysis. Arnold and Press (1989) showed that a gamma distribution, $\text{Gamma}(c_0, d_0)$, i.e., $\pi(\alpha_{ij}) \propto \alpha_{ij}^{c_0-1} e^{-d_0 \alpha_{ij}}$ is the conjugate prior distribution for Pareto distribution. Therefore the posterior distribution of α_{ij} is $\text{Gamma}(c_0 + 1, d_0 +$

$\log(\tau y_{ij})$). Fisher's Information and Jeffreys' prior for a Pareto model are respectively:

$$I(\alpha_{ij}) = -E \left[\frac{\partial^2 \log f(y_{ij} | \alpha_{ij})}{\partial \alpha_{ij}^2} \right] = \frac{1}{\alpha_{ij}^2},$$

and Jeffreys' Prior: $J(\alpha_{ij}) = [I(\alpha_{ij})]^{\frac{1}{2}} \propto \frac{1}{\alpha_{ij}}.$

Jeffreys' prior over canonical parameter α_{ij} is proportional to an improper $Gamma(c_0 = 0, d_0 = 0)$. We consider the link function as the transformation of variables and consequently the posterior density of η_{ij} is

$$\pi(\eta_{ij} | y_{ij}) = C e^{\eta_{ij}((c_0+1)-1)} \exp\{-(d_0 + \log(\tau y_{ij}))e^{\eta_{ij}}\} e^{\eta_{ij}},$$

where $-\infty < \eta_{ij} < \infty$, and C is the normalizing constant. Note that the posterior of η_{ij} follows a log-gamma density with parameters $(c_0 + 1)$ and $(d_0 + \log(\tau y_{ij}))$.

Algorithm Under the full rank assumption of the design matrix X , we can generate samples from the posterior distribution of β using the algorithm as described below. An advantage of this algorithm is that, as long as we know the posterior distribution of α_{ij} for all i and j , we do not need to know the posterior distribution of β to obtain posterior samples of β .

Step 1: Suppose that we are at the r^{th} iteration of the algorithm. Generate samples $\alpha_{ij}^{(r)}$ from $Gamma(c_0 + 1, d_0 + \log(\tau y_{ij}))$ for $j = 1, \dots, n_i, i = 1, \dots, k$ and $r = 1, 2, \dots, N$; where N is the simulation size.

Step 2: Calculate $\eta_{ij}^{(r)} = \log(\alpha_{ij}^{(r)})$ for $j = 1, \dots, n_i, i = 1, \dots, k$ and $r = 1, 2, \dots, N$.

Step 3: Calculate $\beta^{(r)} = (X'X)^{-1}X'\eta^{(r)}$ for $r = 1, 2, \dots, N$.

Step 4: Go to **Step 1** until $r = N$.

Once we have the posterior samples of β , $\{\beta^{(r)} | r = 1, \dots, N\}$, we can do all the necessary inferences on β . Das (2008) showed that in the presence of significant multi-collinearity the regular MCMC technique performs miserably while this *algorithm* works well.

4 Analysis of Extreme Behavior of Patients diagnosed with AD.

Since the definition of heavy drinking differs by sex, we present separate analyses for female and male patients for extreme drinking.

4.1 Analysis of Female Patients

We fit the Pareto regression model for women with AD and present the results in Table 2. We found that the day effect $\{0.1019 \text{ with } 95\% \text{ CI } (0.0747, 0.1248)\}$ is statistically significant. The positive estimate implies that among women the probability of extreme drinking decreases during the treatment (or study) period. The effect of age, $\{0.2728 \text{ with } 95\% \text{ CI } (0.2145, 0.3323)\}$ is also significant and again the positive estimate means that the probability of extreme drinking behavior decreases with age. There was also a significant main effect of *GABRA2* genotype $\{-0.1606 \text{ with } 95\% \text{ CI } (-0.3058, -0.0145)\}$ over extreme drinking behavior among women with AD. The negative estimate shows that female patients with two copies of the low-risk allele had a lower probability of extreme drinking than women with one or two copies of the high risk allele. This effect is shown in Figure 1.a, where the intercept of each curve represents the main effect of each risk group at baseline. At the baseline, women with one or two copies of the high-risk allele had a significantly higher probability of extreme drinking than women in the low-risk group. However, women with one or two copies of high-risk alleles showed a better response to treatment and their probability of extreme drinking decreased faster than women with two copies of low-risk allele.

In Figure 1.b, we present the $P(Y > 30 \mid Y \geq 4)$ along the vertical axis and treatment period along the horizontal axis showing the efficacy of the three different treatments for AD. Bauer et al. (2007) found that TSF was significantly better in treating daily drinking and heavy drinking behavior of patients with AD. However, it is clear from Figure 1.b that among

women CBT was statistically superior to the other two psychotherapies for the treatment of extreme drinking behavior.

4.2 Analysis of Male Patients

We fit the Pareto regression model using the *Monte Carlo algorithm* and present the results for men in Table 3. We found that the day effect {0.1253 with 95% *CI* (0.1133, 0.1372)} was statistically significant. As in women, a positive estimate implies that among men the probability of extreme drinking decreased during the treatment (or study) period. The effect of age {0.2454 with 95% *CI* (0.2108, 0.2807)} was also statistically significant and again the positive estimate means that the probability of extreme drinking behavior decreased with age. The main effect of *GABRA2* genotype, {0.0431 with 95% *CI* (-0.0293, 0.1165)} on extreme drinking behavior among men with AD was not significant. This contrasts with the findings of Bauer et al. (2007), who found that variation in *GABRA2* had a significant effect on the likelihood of both daily drinking and heavy drinking. In figure 2.a, although it appears that high-risk group responded better to therapy than the low-risk group, the effect is not significant {0.0084 with 95% *CI* (-0.0061, 0.0230)}.

In Figure 2.b we present the same $P(Y > 30 | Y \geq 5)$ along the vertical axis and study period along the horizontal axis. Here we present the efficacy of the three different psychotherapy treatments for AD. In Bauer *et al.* (2007), TSF was significantly better in reducing heavy drinking behavior in this AD patient population. However, it is clear from the Figure 2.b that among men MET is significantly superior in reducing extreme drinking to the other two therapies.

5 Discussion

In this paper, we present a detailed analysis of extreme drinking behavior among patients with AD. We present a Pareto regression model with unknown shape parameter to model extreme drinking behavior under a GLM framework. We describe the model for a general data structure and present a Monte Carlo method of fitting the data. We found that the probability of extreme drinking diminished with age and over the treatment period for both sexes. However, the analysis revealed different predictors of extreme drinking behavior in men compared with women. Women with one or two copies of high risk allele of *GABRA2* genotype had a significantly higher probability of extreme drinking than women in the low-risk group. However, high-risk women responded to treatment better and their probability of extreme drinking decreased faster than patients in the low-risk group. Among women, cognitive behavioral therapy was superior to other two treatments, while among men motivational enhancement therapy was superior to the other two therapies in decreasing extreme drinking behavior. The Pareto regression model and the new method of fitting it presented in this paper represent a useful addition to the approaches available for the analysis of data from clinical trials, including those for which genotype is a potential covariate.

Acknowledgement

This material was based upon work partially supported by the National Science Foundation under Grant DMS-0635449 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

- [1] Arnold, B. C., and Press, S. J., (1989). Bayesian estimation and prediction for Pareto data. *J Am Statist Assoc*, **84**, 1079–1084.
- [2] Bauer, L.O., Covault, J., Harel, O., Das, S., Gelernter, J., Anton, R., and Kranzler, H.R., (2007) “Variation in *GABRA2* predicts drinking behavior in project match subjects.” *Alcohol Clin Exp Res* **31**, 1780–1787.
- [3] Behrens, C.N., Lopes, H. F., and Gamerman, D. (2004) “Bayesian analysis of extreme events with threshold estimation” *Stat Modelling*, **4**, 227–244.
- [4] Castellanos, M. E., and Cabras, S., (2007) “A default Bayesian procedure for the generalized Pareto distribution”, *J. Statist. Planning. Inf.* **137**, 473–483.
- [5] Coles, S. G., and Tawn, J. A., (1996) “Modelling extremes of areal rainfall process”, *JRSS - B*, **58**, 329 – 347.
- [6] Covault J, Gelernter J, Hesselbrock V, NellisseryM, Kranzler HR. (2004) “Allelic and haplotypic association of *GABRA2* with alcohol dependence”, *Am J Med Genet B Neuropsychiatr Genet*, **129**, 104-109.
- [7] Das, S., (2008) “Generalized linear models and beyond: An innovative approach from Bayesian perspective.” Ph.D Thesis, University of Connecticut, Storrs.
- [8] Davison, A.C. and Smith,R.L., (1990) “Models exceedances over high thresholds.” (with discussion), *JRSS - B*, **52**, 393–442.
- [9] Gnedenko, B.V. (1943), Sur la distribution limite du terme maximum d’une série aléatoire. *Ann. Math.* **44**, 423–453.

- [10] Kendler, K.S. (2001) “Twin studies of psychiatric illness: an update” *Arch. Gen. Psychiatry*, **58**,1005–1014.
- [11] Lappalainen J, Krupitsky E, RemizovM, Pchelina S, TaraskinaA, Zvartau E, Someberg, L.K., Covault, J., Kranzler, H.R., Hrystal, J.H., Gelernter, J (2005) “Association between alcoholism and gamma-amino butyric acid $\alpha 2$ receptor subtype in a Russian population”, *Alcohol Clin Exp Res*, **29**, 493–498.
- [12] National Institute on Alcohol Abuse and Alcoholism (NIAAA) (2005) “Helping patients who drink too much: A clinician’s guide, 2005 edition”, NIH Publication 05-3769, Bethesda, MD.
- [13] Pierucci-Lagha A, Covault, J., Feinn, R., Nellissery, M., Hernandez-Avila, C., Oncken, C., Morrow AL., and Kranzler, H.R., (2005) “*GABRA2* alleles moderate the subjective effects of alcohol, which are attenuated by finasteride.” *Neuropsychopharmacology*, 30: 1193–1203.
- [14] Pickands, J. (1975) “Statistical inference using extreme order statistics”, *Ann Statist*, **3**, 119–113.
- [15] Project MATCH group (1997 a) “Matching alcoholism treatments to client heterogeneity: Project MATCH posttreatment drinking outcomes.” *J Stud Alcohol*, **58**, 7–29.
- [16] Project MATCH group (1997 b) “Project MATCH secondary a priori hypotheses.”, *Addiction*, 92: 1671–1698.
- [17] Robert. C., and Casella. G., (2005) “Monte Carlo Statistical Methods”, Second Edition, Springer.
- [18] Tancredi, A., Anderson, C., and O’Hagan, A., (2006) “Accounting for threshold uncertainty in extreme value estimation”, *Extremes*, **9**,87–106.

- [19] Tsionas, E. G., (2003) “Pareto regression: a Bayesian analysis.” *Comm Statist: Theor and Method*, **32**, 1213–1225.

Table 1: Data Structure of Extreme drinking

ID	1	2	...	j	...	n_i	\mathbf{x}_i
1	(y_{11}, t_{11})	(y_{12}, t_{12})	...	(y_{1j}, t_{1j})	...	(y_{1n_1}, t_{1n_1})	\mathbf{x}_1
2	(y_{21}, t_{21})	(y_{22}, t_{22})	...	(y_{2j}, t_{2j})	...	(y_{2n_2}, t_{2n_2})	\mathbf{x}_2
\vdots	\vdots	\vdots	...	\vdots	
i	(y_{i1}, t_{i1})	(y_{i2}, t_{i2})	...	(y_{ij}, t_{ij})	...	(y_{in_i}, t_{in_i})	\mathbf{x}_i
\vdots	\vdots	\vdots	...	\vdots	
k	(y_{k1}, t_{k1})	(y_{k2}, t_{k2})	...	(y_{kj}, t_{kj})	...	(y_{kn_k}, t_{kn_k})	\mathbf{x}_k

Table 2: Result for all variables, with estimates and 95% credible interval for corresponding regression coefficients, from the model as describe in section 3 for extreme drinking of female patients

	Estimates	95% CI
Intercept	-1.7356	(-1.9977, -1.4828)
Day (t)	0.1019	(0.0747, 0.1284)
Age	0.2728	(0.2145, 0.3323)
<i>GABRA2</i>	-0.1606	(-0.3058, -0.0145)
<i>GABRA2</i> \times Day	0.0336	(0.0029, 0.0640)
MET	-0.0436	(-0.0834, -0.0042)
CBT	0.1841	(0.1446, 0.2245)

Table 3: Result for all variables, with estimates and 95% credible interval for corresponding regression coefficients, from the model as describe in section 3 for extreme drinking of male patients

	Estimates	95% CI
Intercept	-1.8758	(-2.0151, -1.7337)
Day (t)	0.1253	(0.1133, 0.1372)
Age	0.2454	(0.2107, 0.2807)
<i>GABRA2</i>	0.0431	(-0.0293, 0.1165)
<i>GABRA2</i> \times Day	0.0084	(-0.0061, 0.0230)
MET	0.1131	(0.0912, 0.1363)
CBT	-0.0500	(-0.0723, -0.0277)

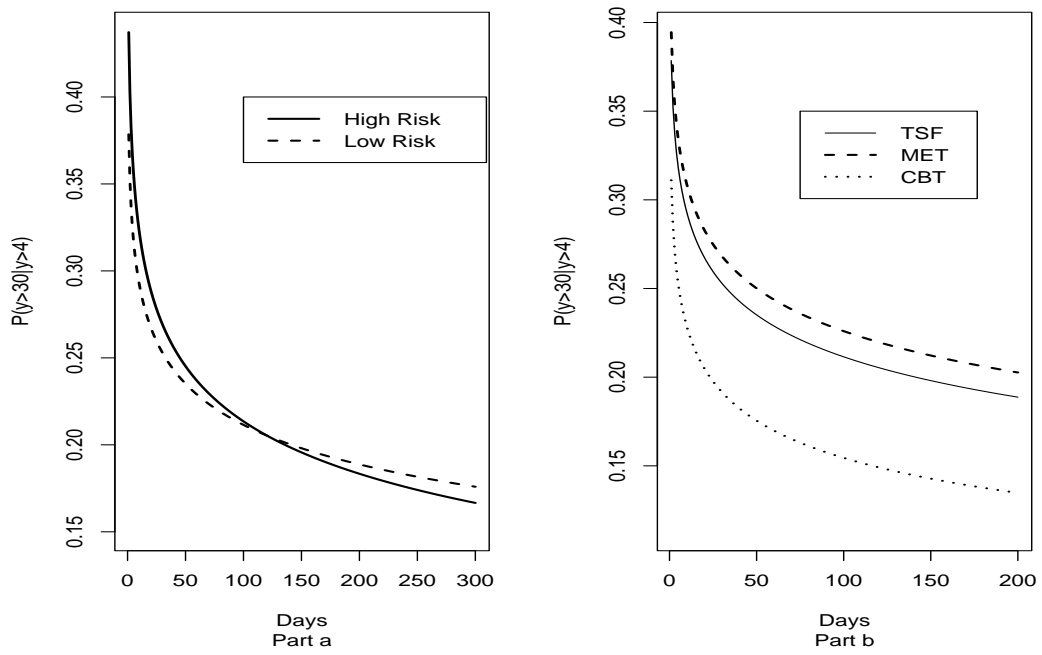


Figure 1: **Part a:** $P(Y > 30 | Y \geq 4)$ for *GABRA2* genotype group for female patients over the study period and **b:** $P(Y > 30 | Y \geq 4)$ for different treatments over the study period

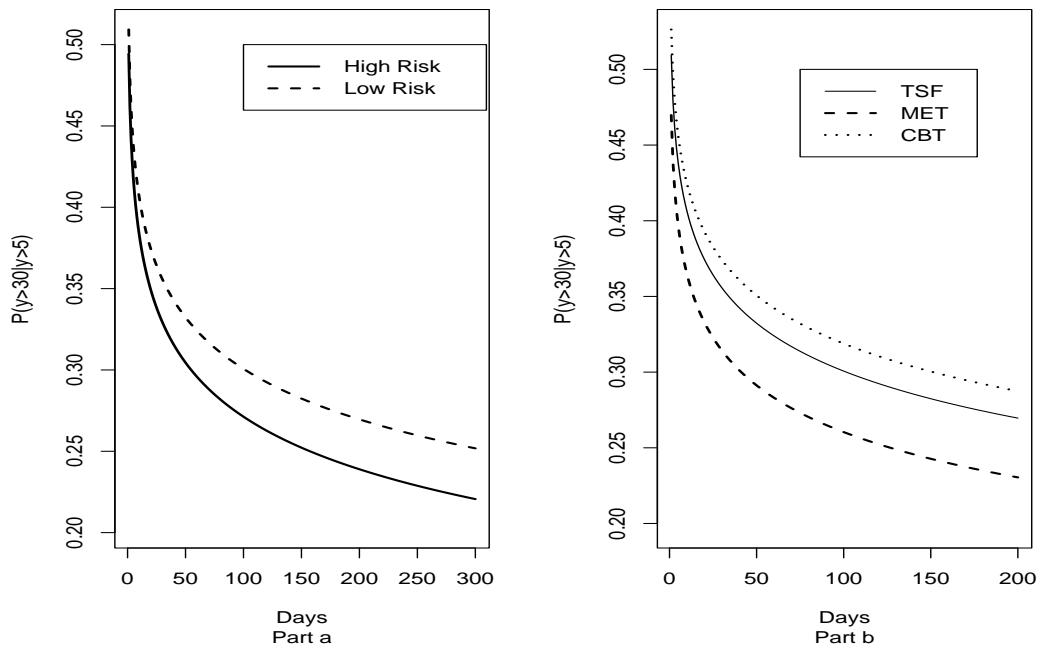


Figure 2: **Part a:** $P(Y > 30 | Y \geq 5)$ for *GABRA2* genotype for male patients over the study period and **Part b:** $P(Y > 30 | Y \geq 5)$ for different treatments over the study period