



Genetic Simulations and Clinical Trials

J. Alan Menius

GlaxoSmithKline Inc.

Outline

- Current Efforts in Clinical Trial Simulation
- The Promise of Pharmacogenetics
- Types of Genetic Information
- Genetic Parameters
- Genetic Simulations
- Multiple Genes and Clinical Variables
- Future Work

Current Efforts in CTS

- Proteomics
- Pharmacokinetics
- Pharmacodynamics
- Candidate Genes

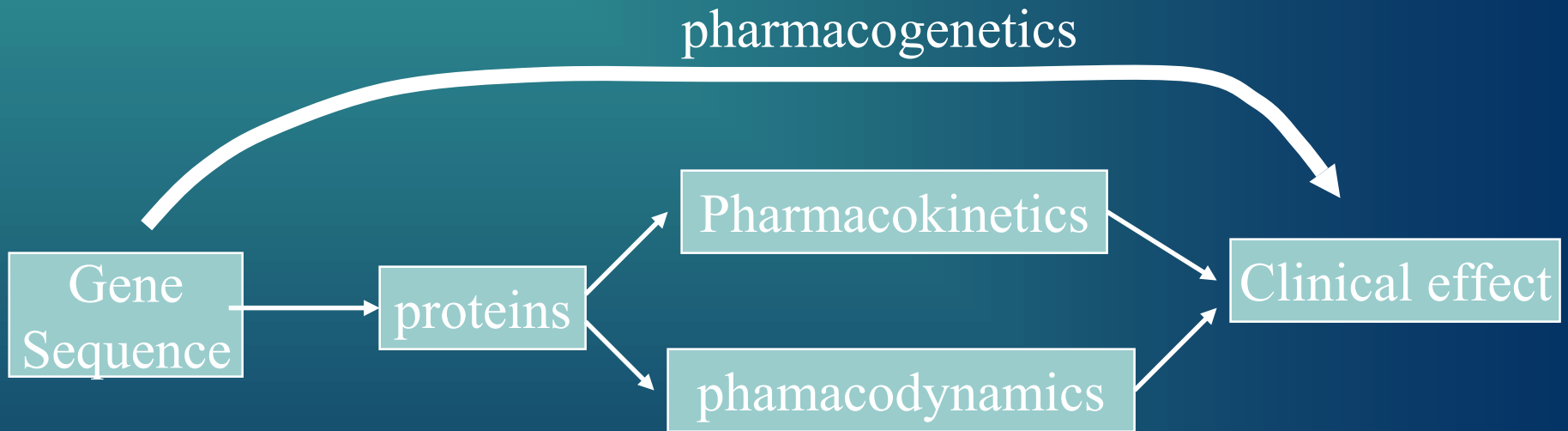
CT Simulation

- Most effort toward pharmacokinetics/ dynamics / phase I
- Compartment models
- Very focused on predicting a particular trial

Is this the correct paradigm for pharmacogenetics

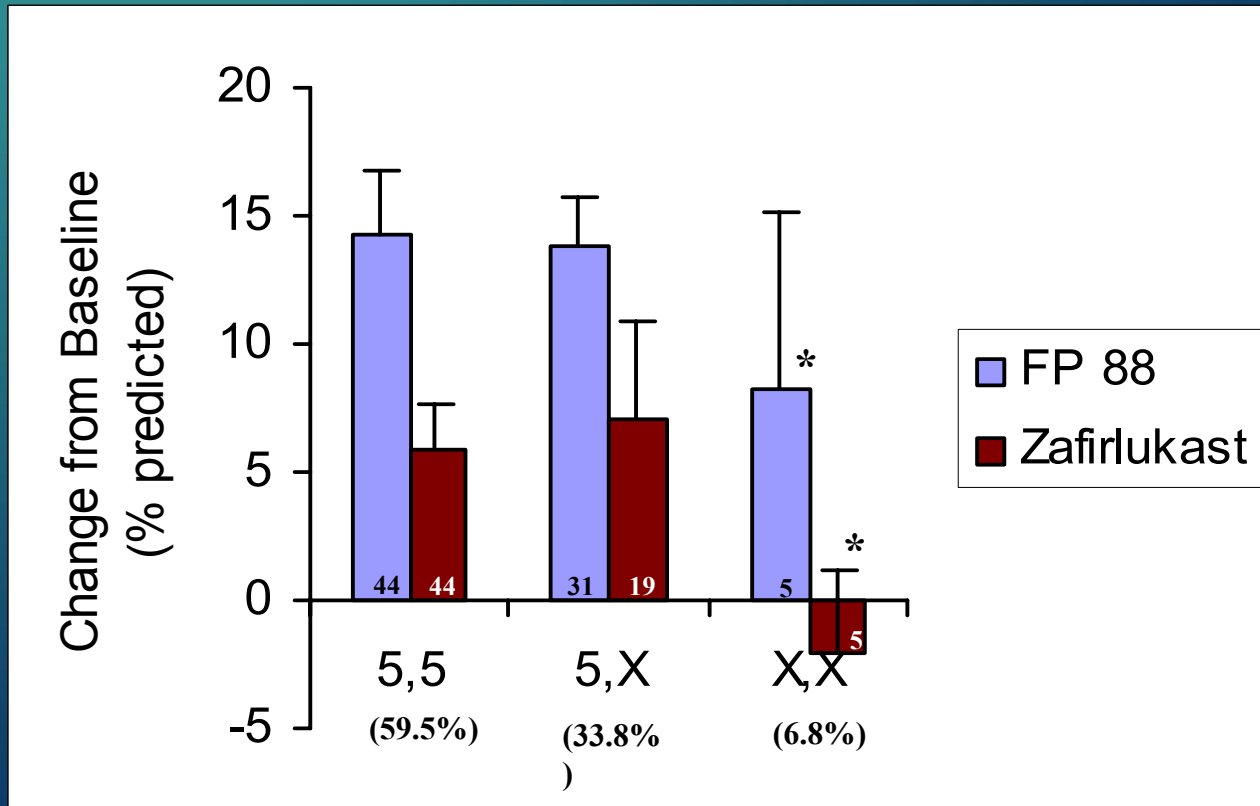
- Most genetics are on candidate genes where function is known
- Genetics added as another parameter in estimating ADME, Toxicity
- We want to know about efficacy and adverse events !

Pharmacogenetics is different !



Assumption: We can use sequence information as indicators of clinical effect

Effect of 5-Lipoxygenase Genotype on Response to Fluticasone and Zafirlukast

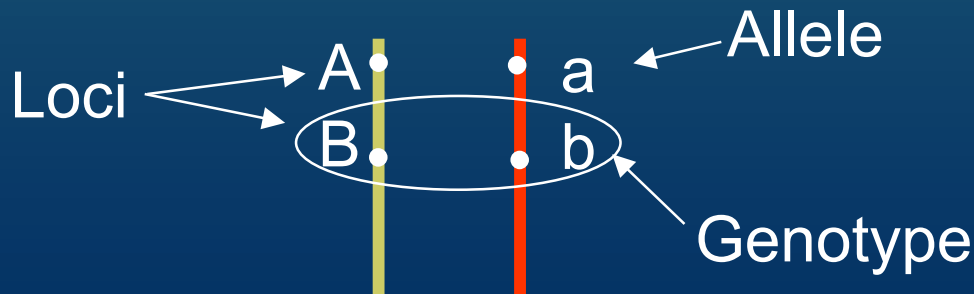


* Not significantly different from baseline

ANOVA indicated a significant drug by genotype interaction

Genetic Terms

- Gene - hereditary unit on chromosome.
- Locus - a position in the genome.
- Marker - position in the genome that can be studied in the lab
- Alleles - variant forms of a gene or position.
- Genotype - pair of alleles at a locus.
- Phenotype - expression of a genotype.



Types of Sequence Information

- Marker
 - Alleles
 - SNPs
 - Repeat polymorphisms
 - Genotypes
- Haplotypes

What is a Candidate Gene?

- Use knowledge of biochemistry, physiology, cell biology, and pharmacology of disease to hypothesize candidate genes
- Evaluate involvement of candidate genes in human disease by statistical genetic analysis
- Identify mutations in candidate genes that cause disease.

Genetic Marker parameters

- Allele frequency
- Linkage disequilibrium
- Penetrance, phenocopy

Hardy Weinberg Equilibrium

- HWE describes a state where:
 - genotype frequencies are constant from generation to generation.
 - genotype frequencies are a product of the allele frequencies.

$$p_{AA} = p_A^2 \quad p_{Aa} = 2p_A p_a \quad p_{aa} = p_a^2$$

a single generation of random mating will restore HWE.

- Forces which can affect HWE include:
 - selection
 - mutation
 - migration
 - nonrandom mating

Hardy Weinberg Equilibrium

- Frequency of A allele in sperm and egg is p_A and frequency of the a allele is $p_a=1-p_A$.
- Genotype frequencies that result from random mating:

		eggs		
		A	a	
sperm	A	AA p_A^2	Aa $p_A p_a$	p_A
	a	Aa $p_A p_a$	aa p_a^2	p_a
		p_A	p_a	

$$p_{AA} = p_A^2$$

$$p_{aa} =$$

$$p_{Aa} = 2p_A p_a$$

$$p_{aa} =$$

Hardy Weinberg Equilibrium

- Allele frequencies of a generation depend on allele frequencies in previous generation, not the genotype frequencies.
- Frequencies of different genotypes produced through random mating depend only on the allele frequencies.
- After 1 generation of random mating the population returns to HWE.

Testing for HWE

	AA	Aa	aa	total
Obs	65	93	42	200
Exp	np_A^2	$2np_Ap_a$	np_a^2	n
	62.72	98.56	38.72	200

- Use χ^2 test to compare observed counts to expected counts where $p_A = (2n_{AA} + n_{Aa})/2n$ and $p_a = (2n_{aa} + n_{Aa})/2n$

$$p_A \sim 0.56 \quad p_a \sim 0.44$$

$$\begin{aligned}\chi_1^2 &= (65-62.72)^2/62.72 + (93-98.56)^2/98.56 + (42-38.72)^2/38.72 \\ &= 0.675 \text{ N.S.}\end{aligned}$$

Linkage Disequilibrium

- Linkage Disequilibrium or Association:
non-independence of alleles at different loci.
- Evolutionary forces such as mutation, drift, and population stratification.
- Close linkage (~30 thousand base pairs) between markers.

Linkage Disequilibrium

Let A be marker locus with alleles A and a having frequencies p_A and p_a










Let B be a marker locus with alleles B and b having frequencies p_B and p_b .

- Linkage disequilibrium between these loci is given by









$$D_{AB} = P(AB) - P(A)P(B)$$

- Linkage disequilibrium indicates that loci are dependent.

Example

		Locus A			Total
		AA	Aa	aa	
Locus B	BB				
	Bb				
	bb				
Total					

Linkage Equilibrium
Uncorrelated Loci

		Locus A			Total
		AA	Aa	aa	
Locus B	BB				
	Bb				
	bb				
Total					

Linkage Disequilibrium
Correlated Loci

Circles represent frequency of multilocus genotype.

Testing for Linkage Disequilibrium

- $D_{AB} = P(AB) - P(A)P(B)$ is linkage disequilibrium
- Can estimate linkage disequilibrium using a type of correlation coefficient, r^2_{AB} .
- nr^2_{AB} has a χ^2_1 distribution.
- r^2_{AB} can be used when markers show HW disequilibrium
- r^2_{AB} can be compared across loci when there is a constant sample size.

Haplotypes

© 2000 Nature America Inc. • <http://biotech.nature.com>

ANALYSIS

Research suggests importance of haplotypes over SNPs

Identifying drug response and disease susceptibility in individuals revolves around the identification of single nucleotide polymorphisms (SNPs) or changes to single bases within genes, but two papers published recently support the theory that the grouping and interaction of several SNPs in haplotypes may be more important. The commercial implications are significant as the development of reliable, cost-effective tests to predict an individual's reaction to a drug or risk of disease will be considerably more complex if haplotypes need to be determined rather than single SNPs. However, many companies are reluctant to restrict investigations to either single SNPs, two or three SNPs, or the haplotypes.

In September, Genaisance (New Haven, CT) and collaborators at the University of Cincinnati College of Medicine for the first time correlated an individual's genetic response to salbutamol with the interaction of multiple SNPs within a haplotype, but could not find a correlation with individual SNPs (PNAS 97, 10483–10488, 2000). In the paper, the researchers identified 13 SNPs in the genetic sequence for the beta-2 adrenergic receptor (the target for salbutamol) within cells from normal, non-asthmatic individuals. By combining these data with information on a reference population of 94 individuals and information from asthmatic patients, Genaisance statistically determined (inferred) the possible haplotypes that actually exist in nature. From the 12 haplotypes determined in this manner only four occurred in the majority of asthma patients analysed. In a subsequent clinical trial, both the haplotypes and the response to salbutamol of 121 asthma patients were determined; it was found that responses to the drug correlated with the patients' haplotype, rather than any individual SNP.

In another study, researchers at GenProfile (Berlin) and Harvard University (Cambridge, MA) have identified the haplotypes for the mu opioid receptor gene (*OPRM1*) associated with heroin and cocaine dependence (*Human Molecular Genetics* 9, in press). By sequence analysis of the *OPRM1* gene in all areas of known functional relevance, including regulatory regions, the researchers identified a substantial number of SNPs in the gene within a group of 172 substance-dependent individuals and controls. Fifty-two haplotypes

were identified by sequencing and classified into two functionally related groups. Five haplotypes were found significantly more frequently in those people with substance dependence than in controls.

Genaisance CEO Gualberto Ruano believes that the haplotype approach may be essential in recognizing genetic factors that create a tendency towards a particular disease or pharmacogenetic phenotype. For example, there may be three SNPs in a gene (including regulatory region, and all coding

"The commercial pressure is now on [the instrumentation companies] to develop a test that reduces the cost and increases the specificity," of haplotype tests.

and non-coding regions), he says, and each individually may alter the gene or its expression in a phenotypically imperceptible manner. However, the combination of the three SNPs may produce the phenotype. Although "some [single] SNPs are predictive [of phenotype], these are usually termed mutations as they have such strong power [affect on the phenotype]," says Ruano. "We are not looking for mutations, but for a more subtle interaction, and this needs the haplotype approach." Other companies focussed on haplotype approaches include Genset (Paris) and Variagenics (Cambridge, MA).

However, some believe that it is individual SNPs that are more important. Jonathan Rothberg, CEO of CuraGen (New Haven, CT), says that "haplotypes are only rarely better than SNPs." But he concedes that a number of approaches should be used, noting that it is more likely that it is the combination of information on "causative [single] SNPs with good medical records," and rigorous statistical analysis that will help companies prioritise drug development. Caragen has built up a large database of SNPs (SNPs present in the expressed regions of genes), which it applies to the identification of disease genes and drug response.

Some commentators, meanwhile, say the distinction between haplotypes and SNPs is artificial and that it's too early to dismiss the potential of either approach. Denis Grant, senior director of pharmacogenetics at

Orchid BioSciences (Princeton, NJ) points out that "in some cases a single SNP will be responsible for a phenotype, whilst in other cases it will be multiple SNPs, either linked [i.e., grouped in a haplotype] or independent [perhaps on different chromosomes]." Mark Edwards, chief scientific officer of Oxagen (Abingdon, UK) agrees. "It is dangerous to generalize [about SNPs and haplotypes]," particularly in the area of disease-associated genes because just not enough is known yet about these genes. Oxagen is working on the identification of genes associated with type 2 diabetes, inflammatory bowel disease, psoriasis, and asthma.

However, the commercial reality is clear: if a single SNP were found to be the basis for variation in drug response or disease risk, the development of a diagnostic product would be comparatively straightforward because it would require a single assay only. Whereas if a combination of several SNPs was found to correlate with phenotype, development of a diagnostic would be more complex and costly, potentially requiring an assay for each SNP involved. As a result, diagnostics based on single SNP assays are expected to reach the market within the next year or two. One of these comes from Gemini Genomics (Cambridge, UK), a specialist in gene discovery through twin and population genomics. It licensed a SNP in the collagen alpha 1 gene to the British-Norwegian diagnostics company Axis-Shield (Dundee, UK) in April 1999 for production of an osteoporosis predisposition test. Gemini's CEO Patrick Kleyn says the SNP is predictive of a much higher chance of fractures in women with osteoporosis.

However, cheap efficient haplotype tests are some way off. "Instruments and kits are not yet at the right level to make it cheap and quick to haplotype people," says Ruano. "The commercial pressure is now on [the instrumentation companies] to develop a test that reduces the cost and increases the specificity," of haplotype tests.

Difficulties lie in haplotype determination in the first place. Direct determination of haplotype (molecular haplotyping) is only possible by sequencing the two homologous chromosomes independently to reveal which SNPs appear on the same piece of DNA—a laborious process that would be very expensive to perform on a large group of people.

Alternatively, algorithms can be used to infer haplotypes based on analysis of SNPs in patients and information on haplotype fre-

© 2000 Nature America Inc. • <http://biotech.nature.com>

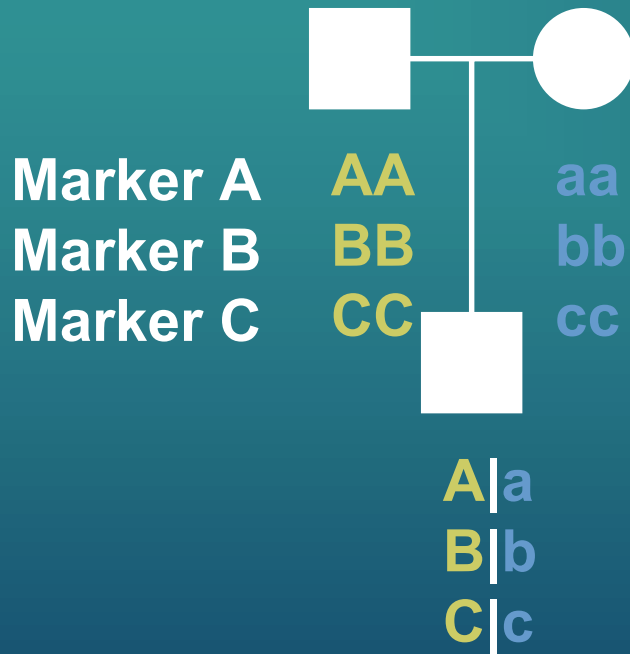
1134

NATURE BIOTECHNOLOGY VOL 18 NOVEMBER 2000 <http://biotech.nature.com>

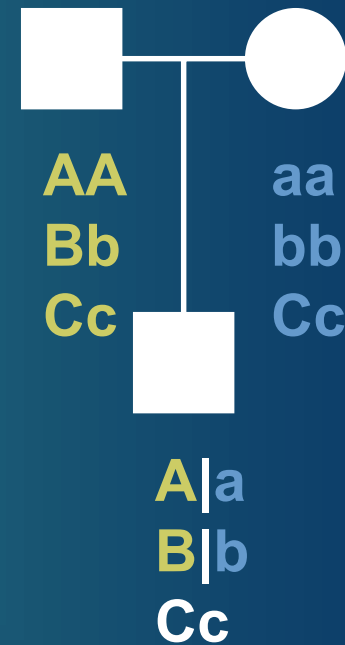
Haplotypes

- The alleles (at different markers) received by an individual from one parent are called a haplotype.
- Molecular biology techniques we currently use cannot identify which alleles came from which parents over long distances.
- Genotype information from parents and grandparents can help elucidate the haplotype.
- Sophisticated statistical techniques can estimate the frequency of haplotypes in a sample without family data (i.e. EM algorithm).

Haplotyping Example



You can distinguish that haplotype ABC came from father and abc from mother.



You can distinguish that haplotype AB came from father and Ab from mother. Marker C is ambiguous

Planning a clinical trial

- Determine effects to be tested
- Utilize 'Pilot' Data
- Power and Sample size computations

Where is the Pilot data for genetics?

- Several large cohorts being studied
- Limited number of genetic markers
- Current answer is to use large numbers to ensure power
- Why not use simulation?

Model Progression

- Monogenic
- Multigenic
- Multigenic + Clinical effects

THE STUDY OF CANDIDATE GENES IN DRUG TRIALS: SAMPLE SIZE CONSIDERATIONS

ROBERT C. ELSTON¹*, RAMANA M. IDURY², LON R. CARDON² AND JAY B. LICHTER^{2†}

¹ Department of Biostatistics and Epidemiology, Rammelkamp Center for Education and Research, MetroHealth Campus, Case Western Reserve University, 2500 MetroHealth Drive, Cleveland, OH 44109, U.S.A.

² AxyS Pharmaceuticals Inc., 11099 North Torrey Pines Road, Suite 160, La Jolla, CA 92037, U.S.A.

SUMMARY

With discovery of an increasing number of candidate genes that may affect inter-individual variability in response to drugs, the design of drug trials that incorporate their study has become relevant. We discuss the determination of sample size for such studies when the number of tests to perform is given, or, alternatively, the number of tests to perform when the sample size is given. In many cases, a uniformly most powerful test does not exist and normal approximations are not sufficiently accurate to determine sample size. We discuss briefly various tests of interest and we give simple examples to illustrate some of the problems that arise. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

In a typical drug trial, one randomizes subjects into dose groups and administers to each subject within a given group the same dose of the drug studied. After a predetermined amount of time, one measures an endpoint response and compares the mean response among groups. We take the variability of response within groups as random and we compare the mean group differences to this random inter-individual variability to evaluate their statistical significance. However, it is now recognized that among major factors determining inter-individual variability in response are variability in the levels of drug metabolizing enzymes (drug pharmacokinetics) or in receptor levels and activity (drug pharmacodynamics). Both of these factors, in turn, may have a major genetic component. There are approximately 100,000 genes in the human genome,¹ many of which may contribute to pharmacokinetic or pharmacodynamic variability. Consequently, there have been several large projects initiated to screen rapidly for polymorphic variants in medically relevant genes² and it is of interest to test whether the polymorphic variants of such candidate genes affect individual response to particular drugs.³ With the rapidly increasing number of

* Correspondence to: Robert C. Elston, Department of Biostatistics and Epidemiology, Rammelkamp Center for Education and Research, MetroHealth Campus, Case Western Reserve University, 2500 MetroHealth Drive, Cleveland, OH 44109, U.S.A.

† Current address: Genset Corporation, 875 Prospect Street, Suite 206, La Jolla, CA 92037, U.S.A.

Contract/grant sponsor: National Institute of General Medical Sciences
Contract/grant number: GM 28356

Contract/grant sponsor: National Center for Research Resources
Contract/grant number: 1 P41 RR03655

A single Candidate Gene

Basic equation used :

$$Z_{1-\alpha/2} - Z_{\beta} = (P_a - P_c) / \sqrt{\frac{P_a(1-P_a)}{N_a} + \frac{P_c(1-P_c)}{N_c}}$$

where P_a and P_c are population frequencies of the allele in cases and controls, respectively, N_a and N_c are total number of *alleles* from cases and controls, and $Z_{1-\alpha/2}$ and Z_{β} are values of the standard normal deviate corresponding to significance level (α) and power ($1-\beta$). This equation was derived by Darvasi and McGinnis (ASHG meeting 1999, abstract 1314).

Reduces to:

(if $N_a = N_c$, then square both sides and cross-multiply)

$$N = (Z_{1-\alpha/2} - Z_{\beta})^2 [P_a(1-P_a) + P_c(1-P_c)] / (P_a - P_c)^2$$

Main issue is the magnitude of effect.

Power and Sample Size for a single Candidate Gene

Tables showing some allele frequencies and genotype effect differences that a case control study of 400 can support at a Power=80%.

p_1	Power	$D_2 - D_1$									
		0.05		0.1		0.2		0.3		0.5	
		N	(O.R.)	N	(O.R.)	N	(O.R.)	N	(O.R.)	N	(O.R.)
0.01	90%	790	(6.32)	320	(12.2)	130	(26.3)	74	(44.5)	30	(103)
	80%	620		252		102		58		24	
0.05	90%	1636	(2.11)	520	(3.35)	174	(6.3)	90	(10.2)	36	(23.2)
	80%	1284		408		138		72		28	
0.1	90%	2590	(1.59)	744	(2.25)	224	(3.86)	110	(6)	40	(13.5)
	80%	2032		584		176		86		30	
0.2	90%	4136	(1.33)	1102	(1.71)	298	(2.67)	136	(4)	44	(9.33)
	80%	3246		864		234		106		34	
0.3	90%	5208	(1.26)	1340	(1.56)	342	(2.33)	148	(3.5)	44	(9.33)
	80%	4088		1052		268		116		34	
0.4	90%	5802	(1.23)	1458	(1.5)	358	(2.25)	148	(3.5)	40	(13.5)
	80%	4554		1144		280		116		30	
0.5	90%	5922	(1.22)	1458	(1.5)	342	(2.33)	136	(4)	32	(nd)
	80%	4648		1144		268		106		26	

Multiple Genes

- Linkage Disequilibrium
- Complex interactions
- Additive effects
- Mixtures

Clinical Factors

Demographic Variables

Age, Sex, Body Weight, etc.

Exposure Variables

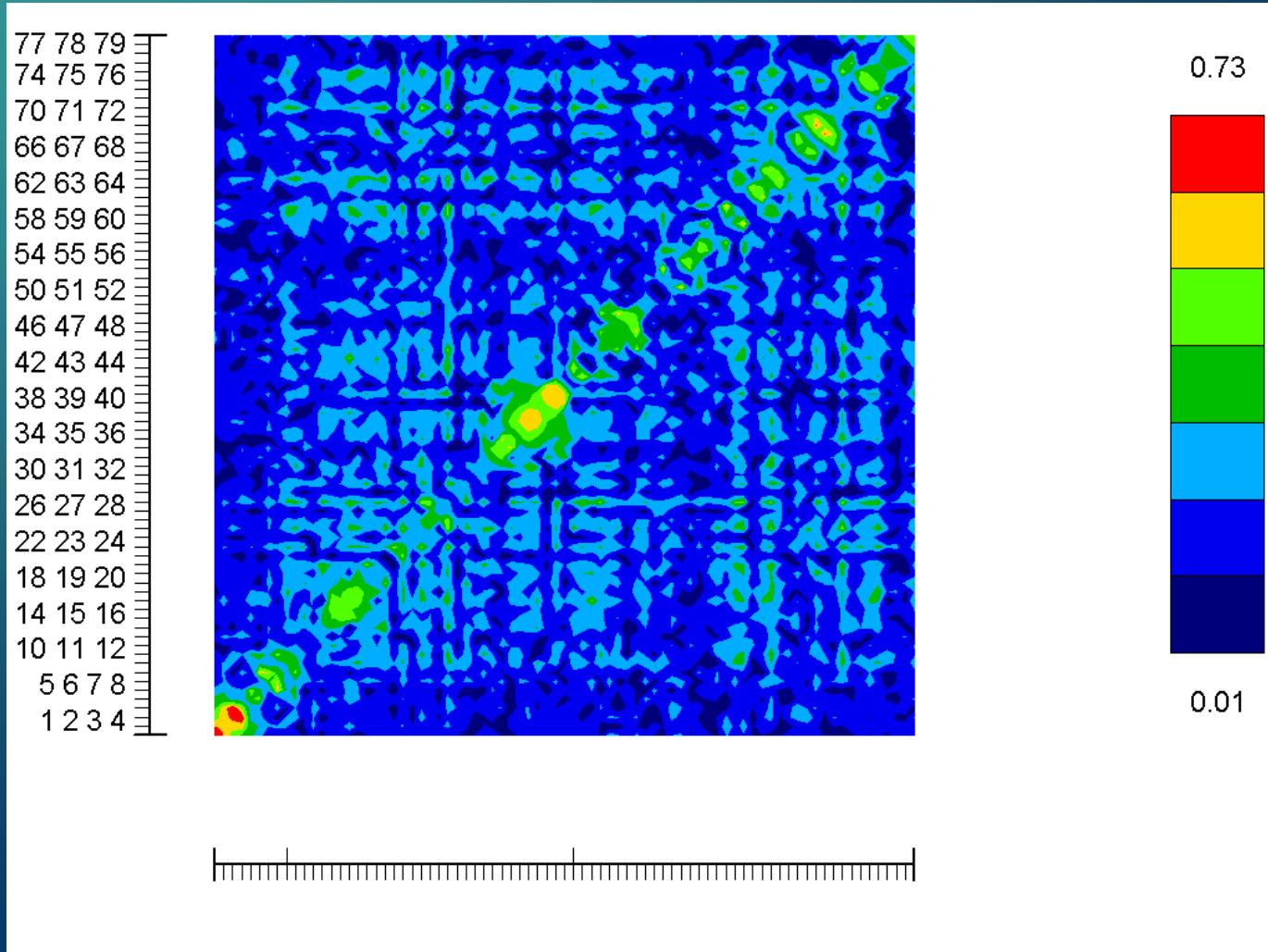
Drugs, Chemicals, Infections, etc.

Continuous, 0/1, multinomial, nominal, missing, etc.

Simulating a population

- Start with n genes and k markers per gene
 - SNPs
 - Satellite Markers
- Hardy Weinberg eq for all markers
- Vary disequilibrium between markers

Linkage Disequilibrium of Genetic Markers



Genetic EXOR effect

	AA	Aa	aa
BB	0	0	1
Bb	0	0	1
bb	1	1	0

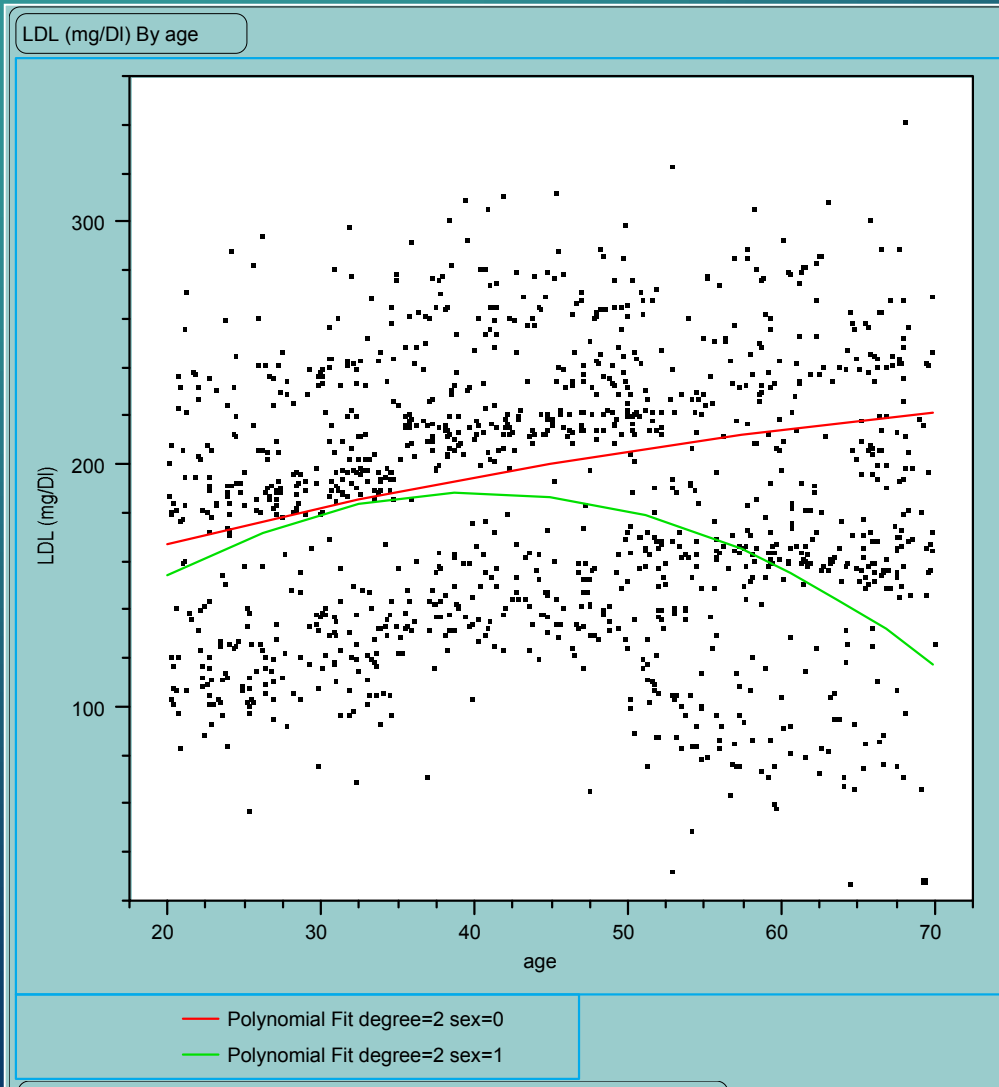
Genetic Interaction

	AA	Aa	aa
BB	0	0	1
Bb	0	0	1
bb	0	0	0

Adding in Clinical Effects

- Utilize population studies for starting parameter estimates and error
- Create genetic effects
 - Additive
 - Interactions
 - mixtures
- Build in correlation among independent variables.

Modeling Environmental Factors



Non-linear sex*age interaction

Building in 'Realism'

- Missing values
- Genes that effect clinical parameters
- Mixtures of genes and clinical parameters

The final model is series of additive functions

$$\text{Predicted Response} = (a(B(c+e1)+e2))$$

Adverse Events

- Small proportion of events
 - 1-10% of population
- Multiple types
- Same possible effects as efficacy models

What are the Questions

- If we assume genetic/environmental interactions exists
 - Can we detect them?
 - Are the stat methods robust?
 - How do we power the study?
 - How will it impact the next study?

In an ideal 'pharmacogenetics' world

- Large phase II - small phase III
- Large phase IV - med phase II - small phase III
- Can we simulate impact of effects from one trial to the next?

Other applications

- Diagnostics
- Pharmacoeconomics
- Commercial / eBusiness

Acknowledgments

GlaxoSmithKline:

Meg Ehm, Mike Mosteller, Linda Surh, Stan Young,
Dmitri Zaykin

UNC-CH:

Marla DeLuca, Gary Koch

References

Cardon L. Testing drug response in the presence of genetic information; sampling issues for clinical trials. *Pharmacogenetics* 10; 503-510 (2000).

Elston R. The study of Candidate Genes in Drug Trials: Sample Size Considerations. *Statist. Med.* 18, 741-751 (1999).