

The Statgen Working Group in IRMACS

Ji-Hyung Shin, Brad McNeney

Simon Fraser University

Department of Statistics and Actuarial Science

Overview of Talk

- Overview of Working Group
- Genetic association studies
- One of our current projects:
 - Improved inference of disease associations with a genetic factor and an independent non-genetic factor in a case-control study.

Overview of Working Group

- Research interests in statistical genetics and genetic epidemiology
<http://stat-db.stat.sfu.ca:8080/statgen/>
- Genetic association studies of collaborators
 - Multilocus Age-dependent Genetic Effects on Type 1 diabetes
 - Genetic risk factors for heart, lung and blood vessel disease
 - Genetic risk factors for non-Hodgkin lymphoma
- Methodological development
 - Finding recombination breakpoints in sequence alignments
 - MITACS biomedical project “Statistical Modeling and Analysis of Complex Traits in Human Populations”
 - Improved inference in genetic association studies
- Implementing methodology in freely available software
 - stepwise, hapassoc, LDheatmap, luca, others in progress

Genetic association studies

- Association studies:
 - Measure a trait of interest and other attributes (“risk factors”) on a sample of individuals from a population.
 - Infer associations between risk factors and the trait from the data.
 - In studies where the trait is disease status, measure association by the odds-ratio (OR)
 - Example OR : odds of lung cancer among smokers relative to odds of lung cancer among non-smokers
- Genetic association studies: association studies with genetic markers among the risk factors

One of our current projects

Improved inference of disease associations with a genetic factor and an independent non-genetic factor in a case-control study.

- Motivating example: associations between type 1 diabetes (T1D) and the glutamate-cysteine ligase catalytic subunit gene (GCLC).
- T1D is a rare autoimmune disease with age-dependent symptoms, immune markers and genetic factors.
- GCLC is involved in synthesis of glutathione, which protects cells from reactive molecules and toxic compounds.
- Exposure to toxic agents (e.g., nitrosamines) is associated with T1D (Dahlquist et al. 1990).
- Association between sensitivity to toxic agents and a variant form or allele, GCLC8, of this gene has been observed (Walsh et al. 2001).

⇒ Hypothesis: GCLC8 may affect the risk of T1D.

Case-Control Study of Type 1 Diabetes

- Investigate the possibility of age-dependent associations between GCLC8 and T1D.
- T1D has age-dependent symptoms and immune markers.
 - Clinical onset is much more abrupt in young patients than in older patients.
 - Young-onset patients have different profile of circulating autoantibodies than older onset patients.
- Age-dependent associations suggest different contributing factors in different age-groups.
- We considered 179 Swedish incident patients with T1D and 186 healthy Swedish controls aged 0–34 years.

Motivation

- TABLE 1 summarizes the GCLC8 status of patients and controls by age, the estimated age-specific T1D risks (OR's) and associated confidence intervals (CI's).

TABLE 1

	age in years									
	[0,7)		[7,14)		[14,21)		[21,28)		[28,35]	
	con ^a	cas ^b	con	cas	con	cas	con	cas	con	cas
GCLC8- ^c	3	12	55	46	29	25	20	25	26	20
GCLC8+ ^d	3	11	23	22	10	8	8	5	9	5
OR ^e	0.92		1.14		0.93		0.50		0.72	
95% CI ^f	(.1–8.4)		(.6–2.3)		(.3–2.7)		(.1–1.8)		(.2–2.5)	

^a “con”, controls; ^b “cas”, T1D patients; ^c “GCLC8-”, individuals with no copies of GCLC8; ^d “GCLC8+”, individuals with at least one copy of GCLC8; ^e “OR”, odds-ratio; ^f “CI”, confidence interval.

- The same data presented assuming GCLC8+ genotype frequencies are constant across ages.

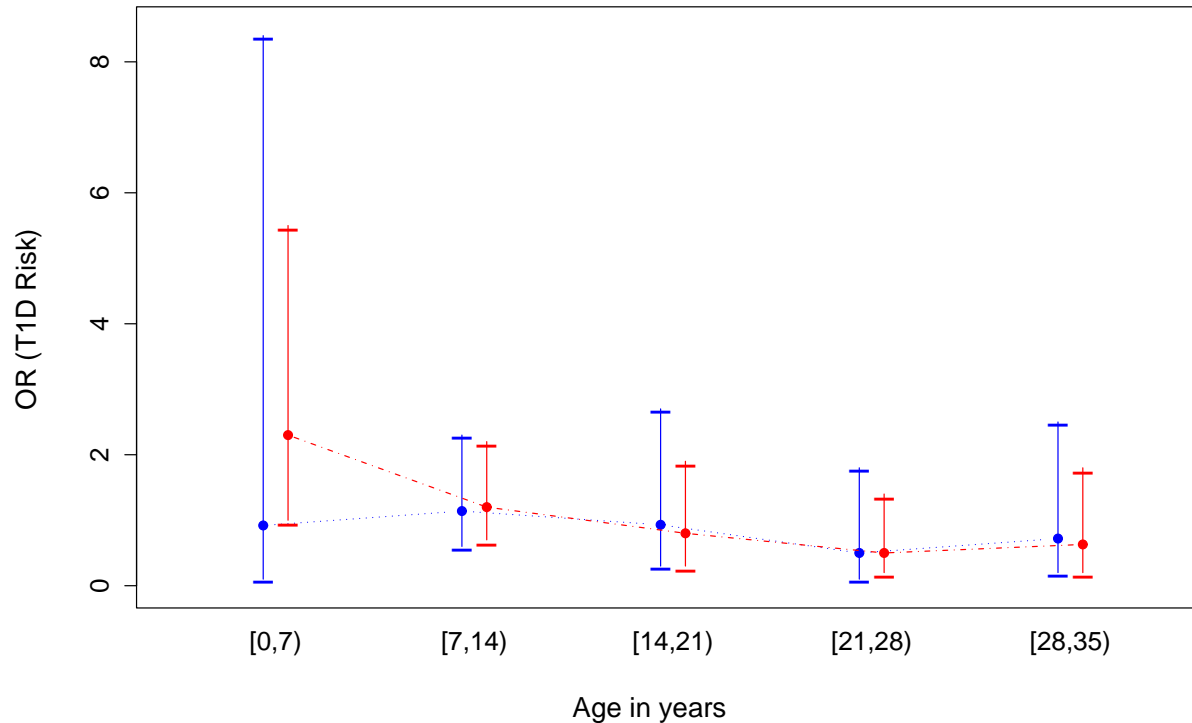
TABLE 2

	con [†]	age in years				
		[0,7)	[7,14)	[14,21)	[21,28)	[28,35]
GCLC8-	133	12	46	25	25	20
GCLC8+	53	11	22	8	5	5
OR	(1.0)	2.30	1.20	0.80	0.50	0.63
95% CI	—	(1.0–5.5)	(0.7–2.2)	(0.3–1.9)	(0.2–1.4)	(0.2–1.8)

[†] “con”, controls of all ages

- Achieved more controls for each age group with no additional cost.

Figure 1: OR's and CI's from TABLE 1 (blue) and TABLE 2 (red).



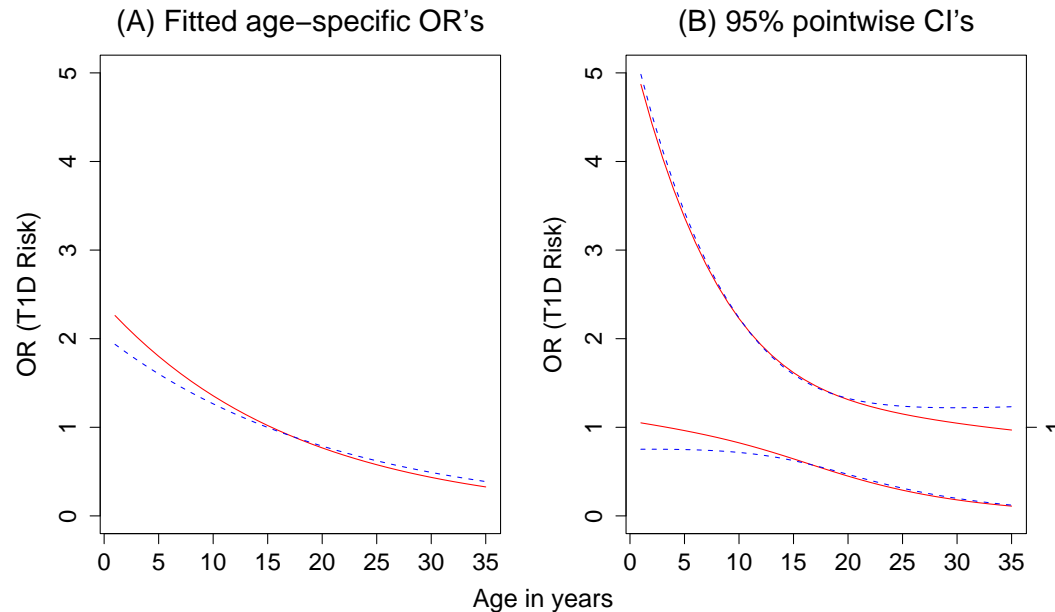
- Clearer pattern of risk decrease by age and narrower CI's from pooling controls: increased precision to detect GCLC8-by-age interaction (age-dependent association).

Proposed Approach and Application

- For TABLE 2, GCLC8+ genotype frequencies were assumed to be constant across ages in the population.
 - Equivalent to assuming GCLC8 status occurs independently of age in the population.
- For a rare disease such as T1D, controls are representative of the general population.
 - Amounts to assuming GCLC8 status occurs independently of age in the controls \Rightarrow controls were pooled.
- We develop a regression method analogous to TABLE 2 that incorporates this assumption that a genetic factor (GCLC8) and non-genetic attribute (age) occur independently in the controls.

Results

Figure 2: OR's (A) and 95% CI's (B) from standard (blue dotted) and new method (red).



- Figure 2A: T1D risk in GCLC8+ individuals decreases with age; comparable risk estimates for both methods.
- Figure 2B: New method has narrower CI's at younger and older ages \Rightarrow increased precision to detect GCLC8-by-age interaction (i.e., age-dependent associations).

Summary of T1D Analysis

- New method incorporates the reasonable assumption that the genetic factor (GCLC8) and non-genetic attribute (age) occur independently in the population.
- The results from the new method ($p=0.02$) but not the standard method ($p=0.10$) indicate GCLC8 associations with T1D are age-dependent.

General Idea

- Improve precision of statistical inference by incorporating reasonable assumptions about the the distribution of genetic (G) and non-genetic (A) risk factors (or covariates)
 - ⇒ LUCA (Likelihood Under Covariate Assumptions).
- Besides G - A independence, can assume ...
 - Genotype frequencies follow Hardy-Weinberg proportions (HWP)
 - explain
 - G - A dependence through a simple regression model for G .

G - A Assumptions

independence

independence+HWP

dependence

Method

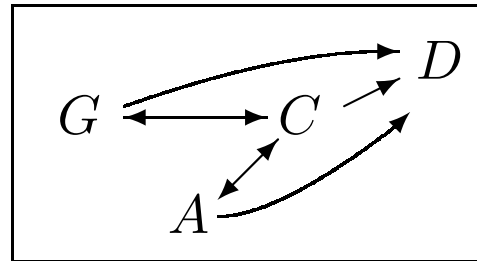
LUCA:Ind

LUCA:Ind+HWP

LUCA:Dep

G - A Dependence

- A key type of dependence that can be accommodated involves a confounding variable for both G and A .
- Suppose G and A are dependent through a third variable C , but are independent given C :



- For instance, in our motivating example, a behavioural risk factor such as smoking (C) could be related to both GCLC8 (G) and age (A), and contribute to T1D risk.
- LUCA:Dep can accommodate the conditional independence of G and A given C .

Statistical Properties of LUCA

- In this talk, focus on power to detect $G \times A$ interactions
 - Simulate genetic factor, non-genetic attributes and disease status on a large cohort
 - Sample 500 affecteds and 500 unaffecteds from cohort
 - Test for presence of $G \times A$ interaction
 - Repeat 10,000 times; estimate of power is the proportion of times a $G \times A$ interaction was detected
- Standard method (logistic regression, or LR) is notorious for low power.

Simulation results: improved power

- LUCA can exploit G - A independence or conditional G - A independence given C to obtain notably more precision than LR.
- Even in limited simulations with $n_{\text{cas}} = n_{\text{con}} = 500, \dots$
 - Under G - A independence, (empirical) power to detect $G \times A$ interaction can be 39% greater for LUCA:Ind than for LR (88 vs 49%).
 - Under conditional G - A independence, power can be 35% greater for LUCA:Dep than for LR (83 vs 48%).
- However, incorporating HWP appears to gain little in precision.

Robustness to assumptions

- Idea of LUCA is to improve power over the standard method (LR) under assumptions such as $G-A$ independence.
- Properties of the approach when such assumptions do not hold are also a concern.
- Does increased power come at the cost of false-positive results when the assumption does not hold?
- Note: LR does not make any assumptions about $G-A$ dependence and so will not lead to increased false-positive results.

Simulation results: robustness

- LUCA:Ind is not robust to non-independence!
 - G - A correlation of 0.025 inflates false-positive rate (type I error) for test of $G \times A$ interaction from 5% to 10%
 - G - A correlation of 0.2 inflates false-positive rate for test of $G \times A$ interaction from 5% to 99.9%.
- Incorrectly assuming HWP has less impact.
 - Inflated false-positive rate from nominal level of 5% to 7.5% for test of $G \times A$ interaction (LUCA:Ind+HWP).

Conclusions

- LUCA:Ind incorporates G - A independence.
- LUCA:Dep incorporates simple G - A dependence, including G - A conditionally independent given C .
- Improved precision and power to detect $G \times A$ interaction for
 - LUCA:Ind under G - A independence. But inference not robust to non-independence
 - LUCA:Dep under conditional G - A independence given C .
- Difficult to detect statistical interaction with case-control data so any improvements in precision are important.
- Use either LUCA:Ind or LUCA:Dep with G - A independence given C to screen for “interesting” interactions to follow up.

Future Work

• Simulations:

- So far have only simulated genetic markers with two variant forms or alleles (SNPs).
- For SNPs, incorporating HWP gains little in precision.
- With genetic markers that are more polymorphic than SNPs, can we gain more precision/power from correctly assuming HWP?

• Software:

- Developing an R-package (`luca`), to be made freely available on CRAN after further testing.