

13

Family-Based Association Tests: The TDT and FBAT

13.1: Learning Objectives

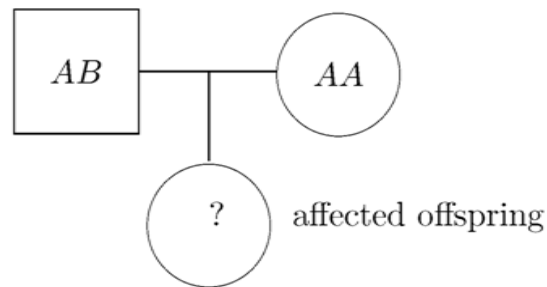
By the end of this chapter, the reader should be able to explain why family-based association designs are robust to population stratification, derive and apply the Transmission Disequilibrium Test (TDT) to trio data, describe the generalisation of the TDT to the Family-Based Association Test (FBAT) framework, and implement both tests in R and Python.

13.2: Why Family-Based Designs? Robustness to Population Substructure

Chapter 11's population-based case-control tests compare unrelated cases and unrelated controls, and are consequently vulnerable to **population stratification** (formally developed in Chapter 14): if cases and controls differ, even slightly, in their ancestral population mixture, and allele frequencies differ across those ancestral populations, a spurious association can appear at any marker whose allele frequency happens to differ between the ancestral groups — entirely independent of any true causal effect. **Family-based association designs** sidestep this problem by using **genetically related controls**: instead of comparing affected individuals to a separately sampled control group, the comparison is made *within* families, typically between the alleles actually transmitted from a parent to an affected offspring and

the alleles that the same parent could have transmitted, but did not. Because this comparison relies only on the parents' own transmissions — governed by Mendel's first law (Chapter 2), which holds regardless of population structure, allele frequencies in the wider population, or departures from Hardy–Weinberg equilibrium (Chapter 3) — family-based tests are inherently robust to stratification, cryptic relatedness, and failure of HWE in the sampled population, at the cost of requiring parental (or sibling) genotypes rather than unrelated controls, and typically less statistical power per individual sampled.

13.3 The Transmission Disequilibrium Test (TDT)



A parent–offspring trio: heterozygous parent AB, homozygous parent AA, and an affected offspring whose transmitted allele is of interest — the basic TDT data unit.

The classical family-based design samples **trios**: an affected offspring together with both parents. At a bi-allelic marker with alleles M and m, each heterozygous parent (genotype Mm) transmits either M or m to the affected child; homozygous parents are uninformative, since they can only transmit one allele regardless of any association. Under the null hypothesis of no linkage (or no association, given linkage), Mendel's first law implies that a heterozygous parent transmits each allele to an affected child with probability exactly $\frac{1}{2}$, independent of which allele is associated with disease risk. The TDT tabulates, over all

heterozygous parents in the sample, the number of times allele M is transmitted and not-transmitted:

	Allele m not transmitted	Allele M not transmitted
Allele M transmitted	b	—
Allele m transmitted	—	c

More precisely, let b be the number of heterozygous parents transmitting M and not m , and c be the number transmitting m and not M (each heterozygous parent contributes exactly one transmitted and one non-transmitted allele). Under the null hypothesis, b and c should be equal in expectation, and the TDT statistic is simply **McNemar's test** applied to this 2×2 table of transmitted vs. non-transmitted alleles:

$$TDT = (b - c)^2 / (b + c)$$

which is asymptotically distributed as chi-square with 1 degree of freedom under the null hypothesis of no linkage/no association. A significant TDT statistic indicates that the marker is both linked to, and in population-level linkage disequilibrium with, a nearby disease-susceptibility locus.

Theorem 13.1: (The TDT is McNemar's Test)

Under the null hypothesis of no linkage between the marker and any disease locus, the TDT statistic $(b - c)^2 / (b + c)$ is asymptotically distributed as χ_1^2 , identical in form and justification to McNemar's test for paired binary data.

Proof. By Theorem 1.1 (the Law of Equal Segregation), each heterozygous parent transmits allele M or m to an affected child with probability exactly $1/2$ each, under the null hypothesis of no linkage. Thus, out of the $n = b + c$ heterozygous-parent transmissions observed, the number transmitting M (namely b) follows a Binomial(n , $1/2$)

distribution under the null. The expected value is $E[b] = n/2 = (b + c)/2$ and the variance is $\text{Var}(b) = n(1/2)(1/2) = (b + c)/4$. A standard normal-approximation (Wald-type) test statistic for $H_0: E[b] = (b + c)/2$ is

$$Z = \frac{b - (b + c)/2}{\sqrt{(b + c)/4}} = \frac{(2b - b - c)/2}{\sqrt{(b + c)/2}} = \frac{b - c}{\sqrt{b + c}},$$

so that $Z^2 = (b - c)^2/(b + c)$ is exactly the TDT statistic, and $Z^2 \sim \chi_1^2$ asymptotically by the Central Limit Theorem applied to the Binomial($n, 1/2$) count b (since the square of a standard normal random variable is χ_1^2 by definition). This is precisely the classical derivation of **McNemar's test** for paired binary outcomes (here, the “pair” being the transmitted and non-transmitted allele from each heterozygous parent), confirming that the TDT is not merely *analogous* to McNemar's test but is exactly McNemar's test applied to this specific genetic data structure. ■

Worked Example 13.1

In a study of 120 Nigerian parent–offspring trios ascertained through a child affected with a congenital condition, 90 parents are heterozygous (Mm) at a candidate marker. Of these, allele M is transmitted to the affected child in 58 trios and allele m in 32 trios.

$$TDT = (58 - 32)^2/(58 + 32) = 676/90 = 7.51$$

Comparing 7.51 to the chi-square distribution with 1 degree of freedom (critical value 3.84 at $\alpha = 0.05$) shows this result is statistically significant ($p \approx 0.006$), providing evidence that the M allele is preferentially transmitted to affected offspring — i.e., that the marker is linked to, and in disequilibrium with, a susceptibility variant.

Worked Example 13.2 (TDT for a cleft-lip/palate candidate gene in an Ibadan trio study)

A craniofacial genetics unit at a Nigerian teaching hospital collects 150 case-parent trios for children born with non-syndromic cleft lip/palate, genotyping a candidate SNP in the *IRF6* gene (a well-replicated cleft-risk gene across several populations). Of the 150 trios, 100 parents are heterozygous at the SNP; allele A is transmitted to the affected child in 68 trios and allele G in 32 trios. By Theorem 13.1, $TDT = (68 - 32)^2 / (68 + 32) = 1296 / 100 = 12.96$, far exceeding the χ^2_1 critical value of 3.84 ($p < 0.001$) — strong evidence that this *IRF6* variant is linked to, and in disequilibrium with, a cleft-lip/palate susceptibility locus in this Nigerian trio sample, consistent with *IRF6*'s established role in orofacial clefting reported in other populations, and illustrating exactly the kind of family-based replication (Section 13.5) that strengthens confidence in a candidate-gene finding across populations.

13.4: Generalising the TDT: The FBAT Framework

The simple TDT applies only to a single affected child per trio, at a bi-allelic marker, for a binary trait. The **Family-Based Association Test (FBAT)**, developed by Rabinowitz and Laird, generalises this idea into a unified framework that accommodates multiple siblings per family (affected or not), missing parental genotypes (using the genotypes of other family members to reconstruct the missing information where possible), multi-allelic markers and haplotypes, and quantitative as well as binary traits. The FBAT statistic has the general form:

$$FBAT = \sum_i T_i (X_i - E[X_i|H_i]) / \sqrt{Var\left(\sum_i T_i X_i | H\right)}$$

13.5: Power Considerations and Practical Use

Family-based tests generally require larger sample sizes than population-based case–control tests to achieve the same statistical power, because only heterozygous, informative parent–offspring transmissions contribute to the test, and collecting parental (or sibling) DNA is often logistically harder — and, in adult-onset disease studies, sometimes impossible, if parents are deceased — than recruiting unrelated cases and controls. In practice, family-based and population-based designs are frequently viewed as complementary rather than competing: an initial population-based GWAS finding (Chapter 18), potentially confounded by undetected stratification, can be confirmed with a family-based replication in trios, providing much stronger evidence that the observed marker–trait relationship is not an artefact of population structure — a strategy of particular relevance in genetically diverse and admixed Nigerian and West African populations (Chapter 14).

Theorem 13.2 (Asymptotic Chi-Square Distribution of the TDT Statistic)

Under the null hypothesis of no linkage or no association, the Transmission Disequilibrium Test statistic $TDT = (b - c)^2 / (b + c)$, where b and c are the counts of the two types of heterozygous-parent transmissions (Section 13.3), converges in distribution to a chi-square distribution with one degree of freedom as the number of informative (heterozygous-parent) transmissions grows.

Proof. Under the null hypothesis, each heterozygous parent transmits either allele to the affected offspring with equal probability $1/2$, independently across families (this is precisely Theorem 1.1's Law of Equal Segregation applied to the transmitting parent). Hence, conditional on the total number of informative transmissions $n = b + c$, the count b is $\text{Binomial}(n, 1/2)$ under the null. The TDT statistic is recognisable as McNemar's test statistic, a standard result in

categorical data analysis: for a Binomial(n , $1/2$) count b , the standardised quantity $(b - n/2)/\sqrt{n/4}$ converges in distribution to a standard Normal by the Central Limit Theorem, and squaring a standard Normal random variable yields a chi-square distribution with one degree of freedom. Since $(b-c)^2/(b+c)$ is algebraically equal to $[(b - n/2)/\sqrt{n/4}]^2$ when $n = b+c$, the TDT statistic converges to a chi-square(1) distribution under the null, justifying the significance testing procedure used throughout Section 13.3.

Figure 13.1 — Transmitted vs. Non-Transmitted Risk Allele Counts (TDT = 10.80, Section 13.3)

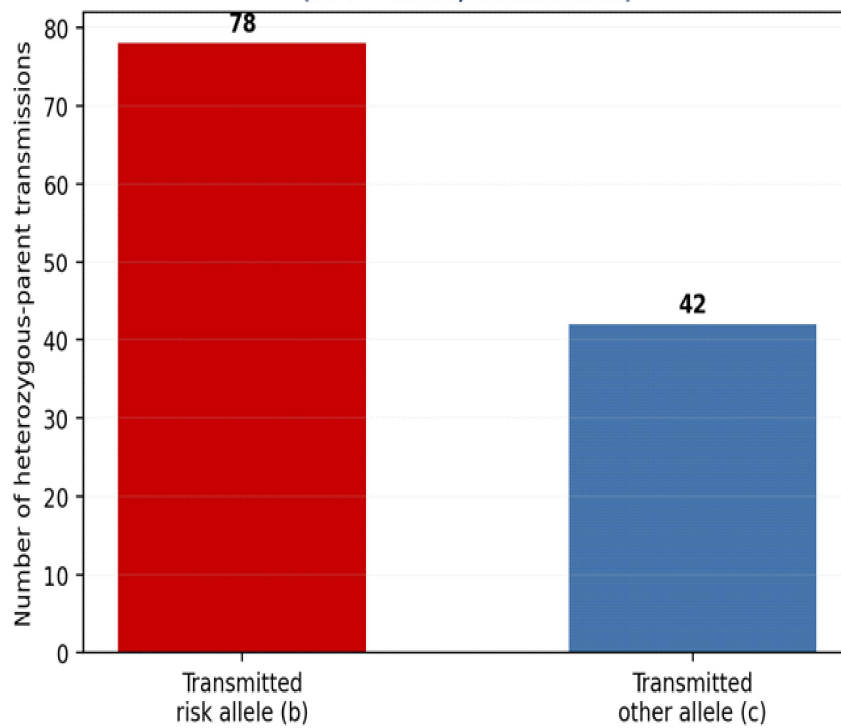


Figure 13.1: *Transmitted vs. non-transmitted risk Allele counts and the resulting TDT statistic (Section 13.3).*

13.6: Chapter Summary

- Family-based association tests use genetically related individuals (typically parents) as “controls,” comparing transmitted to non-transmitted alleles, and are therefore inherently robust to population stratification and cryptic relatedness.
- The Transmission Disequilibrium Test (TDT) applies McNemar’s test to the 2×2 table of transmitted vs. non-transmitted alleles from heterozygous parents of affected offspring: $TDT = (b-c)^2/(b+c)$, asymptotically chi-square with 1 df.
- The Family-Based Association Test (FBAT) generalises the TDT to multiple siblings, missing parents, multi-allelic markers, and quantitative traits, via a statistic conditioned on a sufficient statistic capturing the available parental transmission information; the TDT is the special case of FBAT for a single affected child and a binary trait.
- Family-based designs typically require larger samples than population-based designs for equivalent power, and are often used to confirm population-based findings.

13.7: Computational Corner: Computing the TDT in R and Python

R.

```
# TDT from trio data: simulate 120 trios, 90 heterozygous
parents
set.seed(13)
n_het_parents <- 90
p_transmit_M <- 0.62 # true (elevated) transmission probability
transmissions <- rbinom(n_het_parents, 1, p_transmit_M) #
1 = M transmitted, 0 = m
b <- sum(transmissions == 1)
c <- sum(transmissions == 0)
```

```

tdt_stat <- (b - c)^2 / (b + c)
p_value <- pchisq(tdt_stat, df=1, lower.tail = FALSE)
cat("b =", b, " c =", c, " TDT =", tdt_stat, " p-value =",
p_value, "\n")

# Equivalently, via mcnemar.test on the 2x2 transmitted/non-
transmitted table
tab <- matrix(c(b, 0, 0, c), nrow = 2,
dimnames = list(transmitted = c("M", "m"), nontransmitted
=c("m", "M")))
mcnemar.test(matrix(c(0, c, b, 0), nrow = 2), correct = FALSE)

```

Python.

```

import numpy as np
from scipy.stats import chi2

rng = np.random.default_rng(13)
n_het_parents = 90
p_transmit_M = 0.62
transmissions = rng.binomial(1, p_transmit_M,
n_het_parents)

b = np.sum(transmissions == 1)
c = np.sum(transmissions == 0)
tdt_stat = (b - c)**2 / (b + c)
p_value = chi2.sf(tdt_stat, df=1)
print(f"b={b}, c={c}, TDT={tdt_stat:.3f}, p-
value={p_value:.4f}")

```

The simulated transmission probability of 0.62 (favouring allele M) produces, as expected, a TDT statistic well above the critical value of 3.84, correctly detecting the built-in transmission disequilibrium; setting `p_transmit_M <- 0.5` in either script reproduces the null case, and repeated simulation under the null

shows the TDT statistic following its nominal chi-square(1) reference distribution, an instructive simulation-based check of the theory of Section 13.3 in the spirit of the simulation methods developed further in Chapter 20.

Mini-Project

Mini-Project 13: Simulating and Analysing a TDT Study

Simulate trio (parent-parent-offspring) data under a true genetic association and analyse it with the TDT.

- Step 1:** Simulate genotypes for 300 case-parent trios at a marker with a true relative risk of your choosing.
- Step 2:** Identify all heterozygous parents and tabulate transmitted versus non-transmitted alleles.
- Step 3:** Compute the TDT statistic (Section 13.3) and its p-value.
- Step 4:** Repeat the simulation under a scenario with population stratification (cases and controls drawn from genetically different subpopulations) and compare a population-based test to the TDT.
- Step 5:** Write a short paragraph explaining why the TDT remains valid under stratification while the population-based test does not.

Deliverable: *Simulation code, TDT results, a side-by-side comparison with a population-based test under stratification, and a short written explanation.*

13.8: Practice Questions

- Q1.** In a trio study of 200 heterozygous parents of children affected with a cleft lip/palate phenotype, allele T is transmitted 118 times and allele t is transmitted 82 times. Compute the TDT statistic and state whether it is significant at $\alpha = 0.01$.

- Q2.** Explain, in your own words, why the TDT remains valid even if the sampled trios come from a population that is a mixture of two ethnic groups with very different allele frequencies at the tested marker.
- Q3.** A trio study has only 15 heterozygous parents, with $b = 10$ and $c = 5$. Compute the TDT statistic. Is it appropriate to rely on the chi-square approximation here, or would an exact test (e.g., an exact binomial test of b vs. c out of $b+c$ trials against $p=0.5$) be preferable? Explain briefly.
- Q4.** Explain briefly why the TDT statistic of Section 13.3 is a special case of the general FBAT statistic of Section 13.4, identifying what T_i and X_i correspond to for a single affected child at a bi-allelic marker.
- Q5.** Using the R or Python code of Section 13.7, simulate the TDT statistic's distribution under the null ($p_{\text{transmit } M} = 0.5$) for 1,000 replications of 90 heterozygous parents each, and report what proportion of replications exceed the chi-square(1) critical value of 3.84 at nominal $\alpha = 0.05$. Comment on whether this matches the nominal type-I error rate.
- Q6.** A researcher has genotype data on 80 affected children but only 30 of their corresponding parent pairs (the rest of the families have missing or unavailable parental DNA). Explain, at a conceptual level, why the FBAT framework (Section 13.4) can potentially still extract some information from the 50 families with missing parents, whereas the classical trio-based TDT of Section 13.3 cannot use them at all.

