produce a similar protective effect against reflex sympathetic dystrophy to that of regional sympathetic block immediately before general anaesthesia? Local anaesthetic nerve blocks are usually accompanied by some degree of sympathetic block, and would offer the bonus of postoperative pain relief, and so further reduce the risk of central-nervous-system sensitisation. The complications of local anaesthetic injections into peripheral nerves, especially when an indwelling catheter is used in a neurovascular sheath (eg, during a continuous axillary block), should not be overlooked; indeed, reflex sympathetic dystrophy can be the paradoxical result.

Conclusions

These suggestions for the relief and prevention of reflex sympathetic dystrophy are based on my hypothesis of failed opioid modulation and on the established general principle of avoiding sensitisation of the central nervous system. Although my experiences with treatment of patients with reflex sympathetic dystrophy and of other forms of sympathetically maintained pain tend to support the use of such preventive and therapeutic techniques, controlled studies will clearly be necessary to establish whether or not clinical outcome is significantly improved. Because of the rarity of dystrophic limb disorders, multicentre trials will probably be needed; prospective researchers will first have to address several difficult problems, not least of which will be to reach a satisfactory consensus as to what we really mean by reflex sympathetic dystrophy.

This hypothesis was presented to a symposium sponsored by the Somatosensory Committee of the International Union of Physiological Sciences, held at Guy's Hospital, London, in April, 1991.

REFERENCES

1. Hannington-Kiff JG. Rheumatoid arthritis—interventional treatment with regionally applied drugs and the use of sympathetic modulation: discussion paper. *J R Soc Med* 1990; **83:** 373–76.
2. Di Giulio AM, Yang H-YT, Lutold B, Fratta W, Hong J, Costa E. Characterization of enkephalin-like material extracted from sympathetic ganglia. *Neuropharmacology* 1978; **17:** 989–92.
3. Schultzberg M, Hokfelt T, Terenius L, et al. Enkephalin immunoreactive nerve fibres and cell bodies in sympathetic ganglia of the guinea pig and rat. *Neuroscience* 1979; **4:** 249–70.
4. Helen P, Panula P, Yang H-YT, Hervonen A, Rapoport SI. Location of substance P-, bombesin-gastrin-releasing peptide, [Met^5] enkephalin and [Met^5] enkephalin-Arg^6-Phe^7-like immunoreactivities in adult human sympathetic ganglia. *Neuroscience* 1984; **12:** 907–16.
5. Mays KS, North WC, Schnapp M. Stellate ganglion "blocks" with morphine in sympathetic type pain. *J Neurol Neurosurg Psychiatry* 1981; **44:** 189–90.
6. Arias LM, Bartkowski R, Grossman KL, Schwartzman RJ, Tom CM-T. Sufentanil stellate ganglion injection in the treatment of refractory reflex sympathetic dystrophy. *Reg Anesth* 1989; **14:** 90–92.
7. Janig W. The sympathetic nervous system in pain: physiology and pathophysiology. In: Stanton-Hicks M, ed. Pain and the sympathetic nervous system. Boston: Kluwer, 1990: 17–89.
8. Neely JC. The RAF near-point rule. *Br J Ophthalmol* 1956; **40:** 636–37.
9. Konishi S, Tsunoo A, Otsuka M. Enkephalin as a transmitter for presynaptic inhibition in sympathetic ganglia. *Nature* 1981; **294:** 80–82.
10. Rios L, Jacob JJC. Local inhibition of inflammatory pain by naloxone and its N-methyl quarternary analogue. *Eur J Pharmacol* 1983; **86:** 277–83.
11. Thoren P, Floras JS, Hoffmann P, Seals DR. Endorphins and exercise: physiological mechanisms and clinical implications. *Med Sci Sports Exerc* 1990; **22:** 417–28.
12. Wall PD, Devor M, Inbal FR, et al. Autotomy following peripheral nerve lesions: experimental anaesthesia dolorosa. *Pain* 1980; **7:** 103–13.
13. Roberts WJ. A hypothesis on the physiological basis for causalgia and related pains. *Pain* 1986; **24:** 297–311.
14. An HS, Hawthorne KB, Jackson WT. Reflex sympathetic dystrophy and cigarette smoking. *J Hand Surg* 1988; **13A:** 458–60.
15. Crile GW. The kinetic theory of shock and its prevention through anoci-association (shockless operation). *Lancet* 1913; ii: 7–16.
16. Hannington-Kiff JG. Analgesia during general anaesthesia. *Lancet* 1988; i: 1404–05.

VIEWPOINT

Can meta-analyses be trusted?

SIMON G. THOMPSON STUART J. POCOCK

The enthusiasm for meta-analyses (or overviews) expressed by their proponents is not always shared by the broader medical community. To encourage constructive debate, we adopt a critical perspective on the conduct and interpretation of meta-analysis. We focus particularly on some of the statistical issues, especially heterogeneity between studies, and also on the extrapolation of meta-analysis findings to clinical practice. We conclude that meta-analysis is not an exact statistical science that provides definitive simple answers to complex clinical problems. It is more appropriately viewed as a valuable objective descriptive technique, which often furnishes clear qualitative conclusions about broad treatment policies, but whose quantitative results have to be interpreted cautiously.

Lancet 1991; **338:** 1127–30.

Introduction

A single clinical trial often fails to give clear-cut generalisable results, because of insufficient patient numbers and the particular way the study was done. Meta-analyses (or overviews) use formal statistical techniques to combine the results from similar trials not only to increase numbers but also to generalise conclusions to a more varied range of patients and treatment protocols. The greater objectivity of this approach is a clear advantage over the more subjective narrative review.[1] However, the conviction of some proponents of meta-analysis is often countered by scepticism from the broader medical profession. Two contrasting views are evident:[2] does meta-analysis provide "objective, quantitative methods for combining evidence from separate but similar studies" or merely "statistical tricks which make unjustified assumptions in producing oversimplified generalisations out of a complex of disparate studies"?

In exploring the reasons for such contrasting perceptions of meta-analysis, we will focus on the statistical issues, especially the problems of interpretation arising from clinical heterogeneity (ie, design differences) and statistical heterogeneity. Such a critical appraisal is essential in reaching a balanced view on how meta-analysis findings should be applied in clinical practice. Although we concentrate on clinical trials, similar (and often more serious) issues are relevant to meta-analysis of observational studies.

ADDRESS: **Medical Statistics Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK** (S. G. Thompson, MA, Prof S J. Pocock, PhD). Correspondence to Mr Simon G Thompson.

Which studies to include?

There is a clear distinction between the specific aims of an individual study, assessing, for example, the therapeutic value of one beta-blocking drug in a particular treatment protocol for patients of a certain type after a myocardial infarction, and the broader aims of a meta-analysis—evaluating the efficacy of beta-blockers in general after a heart attack. We want meta-analyses to be relevant to treatment decisions so the studies included must have enough in common for their combined evidence to be interpretable for patients in practice. Given the variety of study designs, treatments, patients, and outcomes a consensus on what to include can be difficult to achieve.

The value of any meta-analysis is totally dependent on lack of bias in its component studies. Hence in a meta-analysis of clinical trials, it is important to restrict inclusion to randomised trials, ideally with intention-to-treat analysis, complete follow-up information, and objective or blinded outcome assessment.[3] In meta-analyses of observational studies, potential biases through confounding, misclassification, or other causes are often more problematic, especially when distinguishing between a small effect and no effect. While it is desirable in principle to down-weight studies of doubtful quality, such weighting in meta-analyses seems too arbitrary, even though quality criteria have been published.[4] For example, is it worse to have poor blinding or incomplete follow-up? The general purpose of meta-analysis—to obtain an objective summary of the evidence available—is better served by defining beforehand acceptable standards for the inclusion of individual studies.

How can a meta-analysis include all relevant studies of an acceptable standard? The collation of all published information is difficult enough[5] but reliance upon published studies alone may distort the results of a meta-analysis because positive studies are more likely to be published than negative ones.[6] Similar selection may influence which abstracts reach publication as full papers[7] or which studies approved by an ethical committee are eventually published.[8] The creation of extensive collaborative groups of investigators has been a notable achievement (eg, the anti-platelet[9] and breast cancer[10] trial groups). Well organised collaborations help to overcome some of the concerns about the inclusion of unpublished studies in a meta-analysis, such as the inadequate quality of interim data and lack of peer review. Computerised registers of published trials[11] and trials in progress[12] should also be encouraged.

Presentation

The first stage in a meta-analysis is a consistent presentation of the principal results from each study. The data shown in the table come from a meta-analysis[13] of nine randomised trials investigating the use of diuretics to prevent pre-eclampsia. For each treatment group, the proportion of patients developing pre-eclampsia is shown. The enormous range of pre-eclampsia risks in the control groups reflects the very different entry criteria and definitions of pre-eclampsia used in the individual trials. To present a meta-analysis, it is necessary to choose a consistent scale for measuring the treatment effects. A simple difference in proportions is often unsuitable since one may expect larger differences in trials of higher risk patients, or in studies with longer follow-up. The alternative is to use risk ratios or odds ratios; where the risks are small (less than 20%) odds ratios and risk ratio analyses will give very similar results. The odds ratio (OR) scale has been used here.

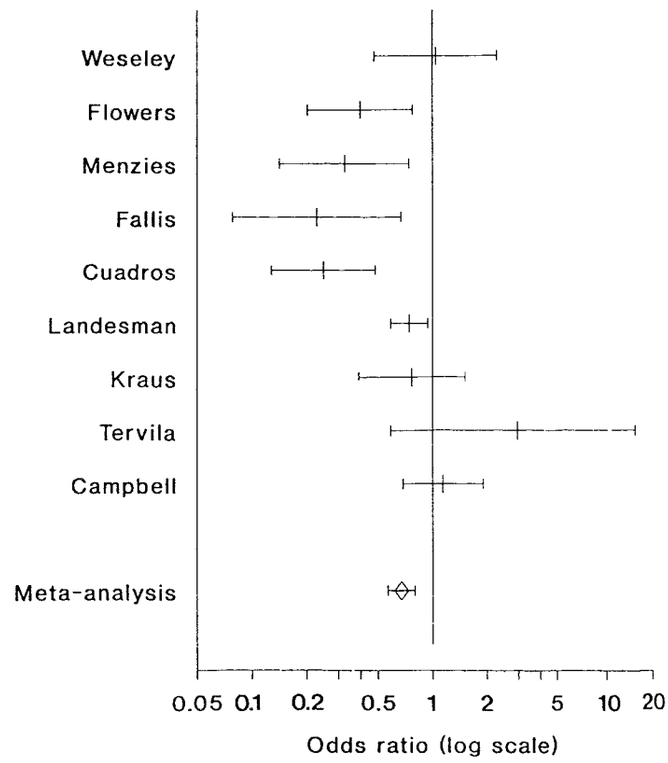
INCIDENCE OF PRE-ECLAMPSIA IN NINE RANDOMISED TRIALS OF DIURETICS, ODDS RATIOS, AND RELATIVE WEIGHTS ASSIGNED IN FIXED-EFFECT METHOD OF META-ANALYSIS

Trial*	Incidence of pre-eclampsia (number of patients)		OR	Weight
	Diuretic	Control		
Weseley	11% (14/131)	10% (14/136)	1·04	5%
Flowers	5% (21/385)	13% (17/134)	0·40	7%
Menzies	25% (14/57)	50% (24/48)	0·33	4%
Fallis	16% (6/38)	45% (18/40)	0·23	3%
Cuadros	1% (12/1011)	5% (35/760)	0·25	7%
Landesman	10% (138/1370)	13% (175/1336)	0·74	55%
Krans	3% (15/506)	4% (20/524)	0·77	7%
Tervila	6% (6/108)	2% (2/103)	2·97	1%
Campbell	42% (65/153)	39% (40/102)	1·14	12%

*Principal author, referenced by Collins et al [13]

The visual display commonly adopted is shown in the figure, which provides estimated ORs for every trial with their 95% confidence intervals (CIs). For example, the OR of 0·40 in the second trial corresponds to an estimated 60% reduction in the odds of pre-eclampsia attributable to the use of diuretics. A logarithmic scale has been chosen so that the CIs are symmetrical. Trials with more events, such as the sixth trial, provide more reliable information and have narrower CIs. Our interpretation therefore has to counter t the misleading visual impression that the least inform ve trials (with the widest CIs) dominate the displa

A 9 , CI that does not include unity corresponds to $p < 0{\cdot}05$ for the treatment comparison. The display therefore serves as a useful reminder that positive (statistically significant) and negative (non-significant) trials are not necessarily inconsistent. For example, the first two trials (one "negative", the other "positive") have CIs which overlap considerably and are therefore not necessarily in conflict. Indeed, the whole concept of classifying trials as



ORs for pre-eclampsia and 95% CIs in nine trials of diuretics.

ORs less than unity represent beneficial effects of diuretics. Meta-analysis based on fixed-effect assumption.

positive or negative according to an arbitrary level of significance is undesirable. However, the figure does leave an impression of real discrepancies between the results of the nine trials. This implies a degree of statistical heterogeneity—which, as we discuss below, makes a meta-analysis difficult to interpret.

Such objective display of results is a valid and instructive presentation which makes no assumptions about the "combinability" of the individual studies. Indeed, careful inspection of such a display often prompts most of the conclusions that will emerge from a more formal statistical analysis of the combined data.

Combining results

The evidence from a collection of trials may be summarised by a test of statistical significance (is there any overall treatment effect?) followed by an estimation of the magnitude of any such effect. The aim of the significance test is to see if the observed departures from no treatment difference add up sufficiently in one direction to a greater extent than could be attributed to chance. Such a test is not trivial since it simultaneously has to deal with several observed treatment differences while taking account of varying trial sizes in an appropriately weighted manner.

The usual procedure for event data (such as the development of pre-eclampsia) is the Mantel-Haenszel test.[14] This test compares, for each trial, the number of events in the active treatment group with the number expected if the treatment had had no effect. The observed and expected numbers are then totalled and a χ^2 test is applied to see if the total of observed events differs from that expected if the treatment had no effect in any of the trials. For the nine trials in the prevention of pre-eclampsia the total number of pre-eclampsia cases in patients on diuretics was 291 as against 343·8 expected ($\chi^2_1 = 21{\cdot}6$, $p < 0{\cdot}0001$). We conclude that the totality of evidence in these trials strongly supports the view that diuretics can prevent some cases of pre-eclampsia.

Since diuretics reduce blood pressure, and since the definitions of pre-eclampsia in these trials depended heavily on blood pressure, this result is not surprising. It is more interesting to estimate the extent to which diuretics prevented pre-eclampsia—indeed, in general, the size of an effect (with its CI) is more informative than a test of whether an effect exists.[15] In the absence of any real differences between the treatment effects observed in each trial, beyond those which can be ascribed to chance, a "fixed-effect" method is appropriate. The assumption is that the underlying true treatment effect in each trial is the same. This single true treatment effect is estimated as a weighted average of the individual trials' estimates. Greater weight is given to trials with narrow CIs; the appropriate weighting is inversely proportional to the variances of the estimates.[16]

Several fixed-effect estimation procedures are available. Woolf's method[17] produces a weighted average of log ORs, weighting being inversely proportional to variance. By anti-logging the results, we can derive the overall OR for the nine diuretics trials as 0·67 (95% CI 0·56–0·80), that is a 33% reduction in odds with a 95% CI of 20% to 44%. The questionable validity of these estimates is discussed shortly, but first it is informative to see how much weight has been allocated to each trial (last column of the table). Over half the weight goes to the sixth trial, whereas four trials each receive 5% or less of the weight. While it is clear that most weight should go to the sixth trial, it is not necessarily desirable that this one study should so dominate the meta-analysis. The meta-analysis result strongly reflects the design and patient characteristics of this one trial, and may not provide a more generalisable conclusion.

Other fixed-effect methods include logistic regression, the Mantel-Haenszel method, and Peto's method. Logistic regression[18] is equivalent to Woolf's method. The Mantel-Haneszel method weights unlogged ORs approximately inversely according to variance;[14] in most instances the choice between OR and log OR is unimportant. The estimate of an overall OR is quite easy to calculate but a reliable formula for obtaining its CI is complex.[19] Peto's method[13,20] gives an approximate overall OR directly through the calculation of the total observed minus expected events used in the Mantel-Haenszel test. The result will be close to that from Woolf's method except where the overall OR is far from unity or where there is great imbalance between the numbers of patients in active and treatment groups.[21] This will only rarely be a problem in clinical trials but one should be wary of the use of Peto's method as a general tool in other situations such as observational studies. All these fixed-effect methods usually produce very similar results: in the example here the overall OR (and 95% CI) are 0·67 (0·57–0·79) for the Mantel-Haenszel method, 0·66 (0·56–0·79) for Peto's and 0·67 (0·56–0·80) for Woolf's.

The main concern with these fixed-effect methods is not which one to choose but the fact that all of them assume no differences between the underlying true treatment effects in the individual trials. However, any set of studies is inevitably clinically heterogeneous by virtue of differences in study design, patient selection, or treatment policy. In a meta-analysis it is more realistic to believe that the true treatment effects vary to some extent. A formal test of whether the observed treatment effects are more dispersed than can readily be ascribed to chance is provided by a test of statistical heterogeneity.[22] However, such a heterogeneity test requires cautious interpretation since it lacks statistical power—even if there is a modest degree of geninue heterogeneity, the test may well be non-significant statistically. So failure to demonstrate heterogeneity statistically does not mean that the studies are truly homogeneous.

For the nine diuretics trials, Woolf's test of heterogeneity[17] gives $\chi^2_8 = 27{\cdot}3$ ($p < 0{\cdot}001$), which is strong evidence that the true treatment effects are different in the trials. The interpretation of a single weighted average of these treatment effects is thus fraught with difficulty; to which diuretic and to what population of patients does the overall OR apply? Even where statistical evidence of heterogeneity is lacking, the same conceptual difficulty bedevils interpretation because of concern about clinical heterogeneity.

Interpretation

The existence of heterogeneity, detected statistically or not, affects how we interpret meta-analyses. The significance test (eg, Mantel-Haenszel) is valid in examining departure from the global null hypothesis that the treatment had no effect in every trial. If, however, some trials had effective treatments and others did not, the overall test might still be highly significant. Overall significance does not permit a blanket conclusion that the treatment effect existed in all the different trial circumstances.

The calculated overall treatment effect is an estimate of a weighted average of (possibly different) true treatment effects in individuals trials. It is a summary of the overall treatment effect for the patients included in the trials covered by the meta-analysis. The extent to which this estimate applies to future patients, depends on the extent to which the trials encompassed a representative collection of patients and treatments. There is invariably a leap of faith between formal statistical inference (which makes the questionable assumption that the trials include a representative sample from a hypothetical population of patients) and extrapolation to the true population of future patients. The overall treatment effect estimated in a meta-analysis provides a useful guide to some average treatment effect in the trials but it is a naïve

oversimplification to regard it as applying directly to future patients.

With a fixed-effect method, the CI for the overall treatment effect reflects the random variation within each trial but not potential heterogeneity between trials. In terms of extrapolation on future patients, the CI interval is therefore artificially narrow. One formal approach to this problem is the random-effects method,[22] which assumes that the true treatment effects in the different trials are randomly positioned about some central value. The method takes into account both random variation within trials and heterogeneity between them, and when heterogeneity is present the CI will be wider than it would have been on the fixed-effect assumption, thus introducing an appropriate degree of statistical caution. However, the random-effects method is no panacea for heterogeneity. Formal interpretation relies on the peculiar premise that the trials done are representative of some hypothetical population of trials, and on the unrealistic assumption that the heterogeneity between studies can be represented by a single variance. The results are also often strongly dependent on the inclusion or exclusion of small trials, which may themselves reflect publication bias. The random-effects methods may therefore give undue weight to small studies, emphasising poor evidence at the expense of good.

A more useful approach is to focus on possible reasons for heterogeneity. For example, it may be fruitful to attempt to explain the heterogeneity in terms of the characteristics of the studies or the patients included, as has been done in a meta-analysis of prospective studies on serum cholesterol and cancer risk.[23] Such a search will not necessarily be successful because the sources of heterogeneity may be intangible, arising, for example, through unrecognised biases in some studies or publication bias. Also, if such inquiry does reveal possible sources of heterogeneity their interpretation will need to be cautious because the analyses are "post hoc" (ie, inspired by looking at the data).

Conclusion

Meta-analysis has the potential to remove idiosyncrasy from the evaluation of medical issues but it is unrealistic to imagine that it will produce simple statistical answers to complex clinical problems. A meta-analysis may provide conclusions about a treatment effect that could not be drawn from individual trials because of small numbers. It may provide evidence about a class of drugs or treatments which allow a general qualitative conclusion to be drawn. Its results are therefore directly relevant to the formulation of broad medical policies. Meta-analysis cannot tell clinicians how to treat an individual patient but it can provide information that helps decision-making.

The problems we have discussed make it unreasonable to interpret simplistically the numerical result a meta-analysis yields. Meta-analysis is a most valuable objective descriptive technique which often provides clear-cut qualitative conclusions. Quantitative conclusions require more care and must take into account the practical relevance of the individual studies making up the meta-analysis and the clinical heterogeneity between them.

REFERENCES

1. Spector TD, Thompson SG. The potential and limitations of meta-analysis. *J Epidemiol Comm Health* 1991; **45:** 89–92.
2. Pocock SJ, Thompson SG. The role of meta-analyses in clinical and epidemiological research. In: Marmot M, Elliott P, eds. Coronary heart disease epidemiology: from aetiology to public health. Oxford: Oxford University Press, 1991.
3. Peto R. Why do we need systematic overviews of randomized trials? *Stat Med* 1987; **6:** 233–40.
4. Chalmers TC, Smith H, Blackburn B, et al. A method for assessing the quality of a randomized controlled trial. *Controlled Clin Trials* 1981; **2:** 31–49.
5. Dickersin K, Hewitt P, Mutch L, Chalmers I, Chalmers TC. Perusing the literature: comparison of medline searching with a perinatal clinical trials database. *Controlled Clin Trials* 1985; **6:** 306–17.
6. Dickersin K, Chan SS, Chalmers TC, Sacks HS, Smith H. Publication bias and clinical trials. *Controlled Clin Trials* 1987; **8:** 343–53.
7. Chalmers I, Adams M, Dickersin K, et al. A cohort study of summary reports of controlled trials. *JAMA* 1990; **263:** 1401–05.
8. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991; **337:** 867–72.
9. Antiplatelet Trialists' Collaboration. Secondary prevention of vascular disease by prolonged antiplatelet treatment. *Br Med J* 1988; **296:** 320–31.
10. Early Breast Cancer Trialists' Collaborative Group. Treatment of early breast cancer. Oxford: Oxford University Press, 1990.
11. Chalmers I, Enkin M, Keirse MJ, eds. Effective care in pregnancy and childbirth: Vol I. Oxford: Oxford University Press, 1989.
12. Simes RJ. Publication bias: the case for an international registry of clinical trials. *J Clin Oncol* 1986; **4:** 1529–41.
13. Collins R, Yusuf S, Peto R. Overview of randomized trials of diuretics in pregnancy. *Br Med J* 1985; **290:** 17–23.
14. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959; **22:** 719–48.
15. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986; **292:** 746–50.
16. Armitage P, Berry G. Statistical methods in medical research. Oxford: Blackwell Scientific Publications, 1987: 194–95.
17. Woolf B. On estimating the relation between blood group and disease. *Ann Hum Genet* 1955; **19:** 251–53.
18. Longnecker MP, Berlin JA, Orza MJ, Chalmers TC. A meta-analysis of alcohol consumption in relation to risk of breast cancer. *JAMA* 1988; **260:** 652–56.
19. Robins JM, Breslow N, Greenland S. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large strata limiting models. *Biometrics* 1986; **42:** 311–23.
20. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Stat Med* (in press).
21. Greenland S, Salvan A. Bias in the one-step method for pooling study results. *Stat Med* 1990; **9:** 247–52.
22. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clin Trials* 1986; **7:** 177–88.
23. Law MR, Thompson SG. Low serum cholesterol and the risk of cancer: an analysis of the published prospective studies. *Cancer Causes Control* 1991; **2:** 253–61.

BOOKSHELF

Biological Perspectives on Human Pigmentation

Ashley H. Robins. Cambridge: Cambridge University Press. 1991. Pp 253. £37.50/$69.50. ISBN 0-521365147.

Melanin has probably caused more social injustice than any other molecule in the body, yet few sociologists understand the biology of skin colour and few physicians are conversant with the anthropological, evolutionary, and psychosocial aspects of racial pigmentation. All these topics are crisply and elegantly reviewed in this monograph from a Cape Town pharmacologist. Robins has worked hard to select the essential facts from many disciplines, and the result is a concise yet thorough overview of this important subject.

The early chapters, which cover skin architecture, the structure of the melanocyte, and the biochemical and hormonal control of pigmentation, will be familiar to many physicians. The chapters on the numerous effects of ultraviolet irradiation on the skin (including recent work in photoimmunology) may be less familiar, and the account of natural photoprotection, sunscreens, and skin lightening and darkening agents will be useful to many general