Perspectives in Diabetes

The Discovery of Type 1 Diabetes

Edwin A.M. Gale

The etiological heterogeneity of idiopathic diabetes has been recognized for 25 years, and subdivision into type 1 and type 2 diabetes is fundamental to the way we think about the disease. Review of the literature suggests that the concept of type 1 diabetes as an immunemediated disease emerged rapidly over the period from 1974 to 1976 and showed many of the features of a classic paradigm shift. A few key observations triggered recognition and acceptance of the new paradigm, but the necessary context was provided by scientific developments in areas mainly unrelated to diabetes. The disease paradigm established by 1976 is still widely accepted, and its essential features have been modified only in detail by the revolution in molecular biology that has occurred over the intervening period. Notwithstanding, some of the underlying assumptions remain imprecise, unchallenged, or unconfirmed. Appreciation of the historical origin and subsequent evolution of these fundamental concepts could stimulate critical analysis and help prepare the way for a new paradigm. Diabetes 217–226, 2001

“"The history of modern knowledge is concerned in no small degree with man's attempt to escape from his previous concepts.”

—Himsworth (1)

“"In research, the present devours the past.”

—Medawar

The word “paradigm” has the root meaning “to show side by side” and is used when an ideal or theoretical model is held up against reality. The historian Thomas Kuhn has pointed out that groups of scientists working within an area form loosely interwoven communities with a common working map, or paradigm (1a). As he put it, “A paradigm is what the members of a scientific community share, and, conversely, a scientific community consists of men who share a paradigm.” Scientific communities can be surprisingly resistant to new ideas or data that do not fit the accepted model, and change, when it comes, comes suddenly. Kuhn refers to this as a “paradigm shift.” Paradigm shifts typically occur among small networks of scientists who interact informally and at meetings. At such junctures, thinking unfreezes and new ideas fly around within the group until the best fit is found and competing views resolve themselves into the new paradigm (1a). Recognition and acceptance of the new paradigm by the wider scientific community comes much more slowly.

Our current view of type 1 diabetes as an autoimmune disorder arose from two dramatic reversals of previous opinion, both of which fit well into Kuhn's classic description of a paradigm shift. The first was the discovery of autoimmunity in the 1950s, and the second was the recognition in the 1970s that juvenile diabetes has an autoimmune basis. Both shifts in perception gained the acceptance of a key section of the scientific community within the space of a few months. In the second paradigm shift, the prevailing view—which pictured early- and late-onset diabetes as gradations or variants of the same genetically determined disease process—was rejected. Instead, juvenile diabetes was seen to fit the paradigm of an organ-specific autoimmune disorder. New knowledge and understanding accumulated so explosively that one authority claimed in 1976 that “new cases of juvenile diabetes may be prevented if our present rate of knowledge about the disease continues to grow over the next two or three years at the rate that it has for the past five years” (2). This promise was not fulfilled, however, and the paradigm established 25 years ago has persisted with minor modification right up to the present. This review will trace the developments leading up to the discovery of type 1 diabetes in 1976 and will ask why the paradigm has proved so stable since then.

THE HETEROGENEITY OF DIABETES

Although Harley, a British physician, commented in 1866 that “there are at least two distinct forms of the disease [diabetes] requiring diametrically opposing forms of treatment” (3), the French physician Lancereaux is generally credited with making the distinction between fat and thin diabetes: diabete gras and diabete maigre (4). The distinction between the two forms of diabetes was indeed stark in the preinsulin era. Most children and some adults died of diabetes within months, whereas overweight older patients often survived for years. When insulin became available, those in the first category no longer wasted away, whereas those in the second continued to get by on diet alone. Clinicians sorted their patients according to the triad of age of onset, perceived requirement for insulin, and body mass, just as they do to this day. Many acute clinical observations were based on this difference in phenotype. For example, Joslin noted that the incidence of diabetes in lean individuals was relatively constant in each decade of life, but that diabetes in the obese was
related to age. He attributed the increasing prevalence of diabetes in the 1930s to increasing obesity (5), this observation being made when there was much less diabetes in the population (6). Clinical observations suggesting that diabetes was a heterogeneous condition were, however, of limited value in the absence of biological markers. As Sydney Brenner has remarked, “Progress in science depends on new techniques, new discoveries and new ideas, probably in that order.” This comment accurately characterizes the discovery of type 1 diabetes.

**Insulin-deficient and insulin-insensitive diabetes.** The first test used to distinguish between the two main forms of diabetes was response to insulin. Wilhelm Falta and other investigators in Vienna drew attention to the existence of insulin-sensitive and -resistant forms of diabetes (7). Insulin-sensitive patients readily suppressed urinary excretion of glucose and developed hypoglycemia in response to a few units of insulin, whereas withdrawal of insulin rapidly resulted in glycosuria and ketosis. These features were lacking in insulin-insensitive patients. In the U.K., Himsworth developed a challenge test in which glucose was given by mouth while insulin was injected intravenously. He found that lean young patients had insulin sensitivity equivalent to that of nondiabetic individuals, whereas older overweight patients were markedly insensitive to insulin. Analysis of arteriovenous glucose levels during these tests led him to conclude that insulin insensitivity resided at the level of skeletal muscle. From these simple clinical observations, he concluded that “in insulin-sensitive diabetics the disease is due to deficiency of insulin, whilst in insulin-insensitive patients diabetes mellitus results, not from lack of insulin, but from lack of an unknown factor which renders the body sensitive to insulin” (8).

**Pancreatic and nonpancreatic diabetes.** What might this “unknown factor” be? Despite the success of pancreatic extract in treating diabetes, pancreata from patients with diabetes often appeared normal under the microscope, and this lent weight to the search for extrapancreatic causes of diabetes. In 1927, it was shown that hyperglycemia could be induced in the dog by injections of extract of anterior pituitary (9); and shortly after this, Houssay, aware that acromegaly had recently been reported to cause diabetes, made his classic observation that hypophysectomy improved features of diabetes. In 1951, 1 ml of plasma from each of six patients was injected into alloxan-induced diabetic, hypophysectomized, and adrenalectomized rats, and the fall in blood glucose was measured after one hour; standard curves were constructed by injection of commercial insulin into the same animals. The study appeared to demonstrate the existence of insulin-deficient and non–insulin-deficient forms of diabetes (15). Meanwhile, attempts had been made to measure the insulin content of pancreata removed at autopsy. These culminated in a massive report from Toronto that described the extractable insulin content of pancreata from 64 diabetic and 139 nondiabetic individuals. Insulin content was assessed by the mouse convulsion test used at that time to calibrate batches of commercial insulin. This study demonstrated that insulin was almost undetectable in the pancreata of diabetic patients who died before the age of 20 years, whereas pancreata from individuals over that age contained on average 40–50% as much insulin per gram wet weight as did pancreata from nondiabetic individuals (16).

**THE GENETICIST’S NIGHTMARE**

It takes an act of imagination for the modern reader to conceive of studying genetics in the absence of any gene markers. What geneticists studied was inheritance, and classic geneticists developed highly sophisticated statistical models for analysis of complex data. Harris, writing of diabetes in 1950, was typical of others in wondering if “we were dealing not with a single disorder but several genetically distinct diseases” (17). He noted the high sib-sib correlation in early-onset diabetes, suggesting a genetic basis for the disease. At the same time, he noticed that “cases of late onset and mild diabetes occur not infrequently among the parents and other relatives of the severe juvenile and young adult forms of the disease [and] this evidence is hardly in line with the view that we are dealing with two genetically distinct and separate diseases” (17). He concluded that early onset patients are homozygous for a gene present in heterozygous form in later-onset disease. By 1970, almost every mode of transmission had been proposed for diabetes: autosomal-recessive, autosomal-dominant, juvenile-homozygous/adult-heterozygous,
sex-linked, and multifactorial (18). Simpson (19) had proposed polygenic inheritance, and this was supported by other studies (20,21). In 1966, Rimoin concluded that definitive genetic studies of diabetes would not be possible until the basic defects underlying the disease were elucidated and reliable disease markers became available (18).

THE AUTOIMMUNE PARADIGM

Ehrlich commented in 1900 that the organism “possesses certain contrivances, by means of which the immunity reaction, so easily produced by all kinds of cells, is prevented from acting against the organism’s own elements and so giving rise to autotoxins […] so that one might be justified in speaking of a horror autotoxicus” (22). It is less well-known that he had also predicted that “when the internal regulating contrivances are no longer intact […] great dangers arise. In the explanation of many disease-phenomena, it will in the future be necessary to consider the possible failure of the internal (i.e., immune) regulation” (23). It was not Ehrlich but others who elevated horror autotoxicus to the status of a biological law.

Noting that allergic encephalomyelitis could be induced in rabbits by injection of brain extracts from their own species, Witebsky and Rose demonstrated in 1957 that antibodies could be generated in the same species against crude extracts of thyroid and against purified thyroglobulin, a protein specific to the thyroid. The identification of isoantibodies directed against tissue from other animals of the same species was soon followed by the demonstration that antibodies could produce autoantibodies against their own thyroid tissue when this was reinjected together with adjuvant. Histological changes appeared in the thyroid after injections of extract or of purified protein (24), and a surgical collaborator pointed out the resemblance to Hashimoto’s thyroiditis (25). The group went on to demonstrate autoantibody formation in patients with this disease, but by the time their report appeared, Roitt and Doniach had also described human thyroid autoantibodies and suggested that “destruction of the thyroid in Hashimoto’s disease results from progressive interaction of thyroglobulin in the gland with the autoantibody present in the patient’s circulation” (26). Within months, yet another group had confirmed that antibodies were present in almost all cases of Hashimoto’s disease and had demonstrated complement-fixing antibodies directed against the adrenal in two patients with Addison’s disease (27). These patients also had thyroid antibodies, and because lymphocytic infiltration of the thyroid had long been recognized in cases of idiopathic Addison’s (28), an autoimmune basis was suggested for both diseases.

The paradigm shift was abrupt. Modern immunology emerged from its chrysalis in 1957, the year in which Macfarlane Burnet summoned the staff at the Walter and Eliza Hall Institute and bluntly informed them that the Institute was switching from virology to immunology. The implication was clear: get into immunology or get out (29). In Burnet’s own words “instead of being concerned primarily with the phenomena of immunity against microbial infection, immunologists are primarily interested today in the way in which the body maintains its genetic and biochemical integrity and in possible ways in which this mechanism can be circumvented in the interests of therapy […] or may by its spontaneous malfunctioning give rise to serious disease” (30). Other clinical disorders were rapidly fitted to the autoimmune model, and in 1963, Mackay and Burnet published a book on autoimmune disease that showed the paradigm essentially in place (31).

Autoantibodies were commonly detected in the 1960s by complement fixation, tanned red cell hemagglutination, or indirect immunofluorescence (32). The latter is a two-step technique. In step one, the serum to be tested is washed over a frozen section of the target tissue so that any antibodies present in the serum will bind to that tissue. In the second step, the tissue section is exposed to antibodies to human immunoglobulin labeled with fluorescein isocyanate. The antibody to be tested is thus sandwiched between the tissue antigen and the labeled anti-immunoglobulin antibody, and the antigen-antibody complexes will fluoresce under the microscope at the appropriate wavelength of light. This widely used method was later used to detect islet cell antibodies (ICAs), but only after many unsuccessful attempts to do so. The reasons for this will become apparent.

Another significant step in the story was the demonstration by Berson and Yalow that insulin-binding globulin was present in the serum of insulin-treated diabetic patients (33). The concept that insulin could itself elicit an immune response suggested immediately that this response could in some way contribute to the development of diabetes, although early models of transient diabetes were clearly due simply to elimination of circulating insulin by anti-insulin serum (34). Injection of thyroid (24) or adrenal (35) extracts had been shown to produce lymphocytic infiltration of the target organ in experimental animals, and this was soon followed by recognition that islet lesions resembling insulinitis could be induced by similar means (36,37). The experiments had to be performed with extracts of whole pancreata, since islet-separation procedures had yet to be developed (38); similar experimental lesions were reported after injection of islet material in 1974 (39). Immunization by injection of target tissue was abandoned when animal models of spontaneous autoimmune diabetes became available (see below), but immunization is still the basis of experimental models used to study multiple sclerosis or rheumatoid arthritis.

Diabetes was already under consideration as an autoimmune disease by the early 1960s, although it was not included as such by Mackay and Burnet in their 1963 book. As Mackay later put it, “The climate of opinion on autoimmunity in 1963 seemed too inhibitory to consider the inclusion of that disease!” (40).

THE ROAD TO AUTOIMMUNE DIABETES

There would in fact have been little justification for listing juvenile diabetes as an autoimmune disorder in 1963. In 1957, Witebsky et al. (24), with Koch’s postulates in mind, had suggested these criteria for autoimmunity: 1) the direct demonstration of free, circulating antibodies that are active at body temperature or of cell-bound antibodies by indirect means; 2) the recognition of the specific antigen against which this antibody is directed; 3) the production of antibodies against the same antigen in experimental animals; 4) the appearance of pathological changes in the corresponding tissues of an actively sensitized experimental animal that are basically similar to those in the human disease.” By 1963, none of these had been described in diabetes. Recognition of these features and of their clinical significance took a further 10 years to accomplish, and many strands of observation fused together.
together in the new paradigm. In approximate (but overlapping) chronological sequence these were:

- induction of experimental immune-mediated islet lesions in animals, and the rediscovery of insulitis in man
- observations showing an excess of organ-specific antibodies against nonpancreatic tissues in insulin-treated but not non–insulin-treated diabetes
- demonstration of cell-mediated autoimmunity directed against islet tissue
- new evidence for a viral etiology for juvenile diabetes
- further clinical evidence of the heterogeneity of diabetes
- identification of immune-response genes in the mouse
- demonstration that juvenile diabetes was associated with specific HLA alleles
- identification, after many previous unsuccessful attempts, of ICAs
- development of animal models of immune-mediated diabetes

In due course, these developments, outlined in the remainder of this review, prefaced the reintroduction of the terms type 1 and type 2 diabetes in 1976.

**The rediscovery of insulitis.** Lymphocytic infiltration of the pancreatic islets was described as an incidental finding by pathologists at the start of the twentieth century, and the term “insulitis” was first used by von Meyenberg in 1940 (41). The lesion escaped greater notice because it is transient, not always obvious, and is found only in the islets of patients with early-onset autoimmune diabetes who die soon after diagnosis. In 1925, Warren examined six childhood-onset cases who died within 2 years of diagnosis because “if there were a recognizable anatomic basis for diabetes, such cases as these should show it best” and concluded that “the absence of demonstrable injury indicates that the disease in children is not ordinarily due to destruction of island tissue” (42). He did encounter “classic” instances of insulitis—for example, in a 6-year-old girl with previously undiagnosed diabetes who died in a coma (43)—but regarded this as a rare exception. His overall conclusion was that “there is no distinctive lesion in the young, uncomplicated cases of diabetes, as might be expected if there were one definite causal agent giving rise to the disease” (44).

LeCompte regarded insulitis as a “rare but possibly significant lesion.” In his account of four cases, he speculated that the lesion might have an infectious or toxic origin and stated that “an antigen-antibody reaction must also be considered” (45). In 1965, Gepts, who had worked with LeCompte, published a careful analysis of pancreata from patients with juvenile diabetes, 22 of whom had died within 6 months of diagnosis (46). He noted insulitis in 15—a much higher proportion than in previous series—and commented that “it seems probable that in the pancreas of acute diabetics we had the opportunity to catch the final stages of a process which has been going on for an indefinite time, perhaps from birth on.” Noting that inflammatory infiltrates resembling insulitis had been induced in rats receiving anti-insulin serum (37) and in cows receiving bovine and porcine insulin in Freund’s adjuvant (36), he tentatively speculated that the islet lesions might have “an immunological origin.”

**Overlap with other autoimmune disease.** As the autoimmune paradigm became established, the overlap between thyroid, gastric, and adrenal autoimmunity rapidly became evident. Thyroid and gastric antibodies were repeatedly noted in the serum of patients with diabetes in early studies, and it soon became apparent that most of those with antibodies were on insulin. Ungar et al. (47) speculated in 1967 as to an autoimmune basis for diabetes in the light of the high prevalence of thyroid and gastric antibodies and Gepts’s report of insulitis and commented that “the most direct evidence for autoimmunity being concerned in the causation of diabetes mellitus would be the regular demonstration in the serum of untreated diabetics of antibody to either islet tissue or insulin, but such evidence is lacking.” Three years later, Irvine et al. (48) commented that “there seems to be a disorder of the immunological system related to insulin-dependent diabetes with respect to the formation of autoantibodies and the occurrence of organ-specific autoimmune disease,” but did not suggest that diabetes was itself immune-mediated. Further provocative observations came in a series of reports from Nerup showing that the prevalence of Addison’s disease and of adrenal antibodies was increased in insulin-dependent diabetes and, conversely, that diabetes developed more commonly in patients with Addison’s (49).

The limiting factor in all of these studies was the failure to demonstrate circulating autoantibodies in diabetes. Early attempts were made, and insulin autoantibodies were reported as early as 1963 in diabetic patients who had not received insulin (50). This report was clearly spurious, because no difference in antibody levels was found between the insulin naive and the insulin-treated at a time when commercial insulin preparations were highly impure. Shortly after this, islet immunofluorescence was reported in 8 of 14 insulin treated patients and 1 of 3 insulin-naive patients (51) and was attributed to insulin antibodies. The pattern of islet staining depicted in this report suggests that this was not an early observation of ICAs. As will be seen, the attempt to identify ICAs by indirect immunofluorescence was made unsuccessfully by investigators including Doniach, Irvine, Mackay, Maclaren, and Nerup, in hundreds of patients and over several years before they were finally demonstrated in 1974.

**Cell-mediated autoimmunity in diabetes.** Although evidence of humoral autoimmunity directed against pancreatic islets was lacking, evidence that cell-mediated immunity was involved came from the work of Nerup, who had previously demonstrated it in Addison’s disease. This was based on the recently introduced leukocyte migration test (52). Mononuclear cells from the subject to be tested were exposed to porcine islet material, spun down into capillary tubes, and incubated on the surface of an agar plate. Under normal circumstances, macrophages will migrate out of the tube, but this process is inhibited if T-cells in the mixture respond to the antigen by production of inhibitory factors. The assay therefore detects the presence of T-cells primed against the antigen tested. These mechanisms were of course unknown at the time, but the test was recognized as a measure of “cellular hypersensitivity,” although many remained skeptical as to its value.

Nerup tested 22 patients ranging from 11 to 75 years of age, 7 of whom had not received insulin. The migration indices of 12 of the 22 patients were reduced, including 5 who were insulin naive. The inhibition could not be achieved with insulin itself, indicating that this was not the antigen. Intracutaneous testing was additionally performed using pig islet material, and 4 of 6 of the diabetic patients with low migration indexes showed a typical delayed-type hypersensitivity.
reaction. Nerup speculated that cellular hypersensitivity was the counterpart of the infiltration seen in insulinitis, and that cell-mediated immunity could therefore play an important part in the pathogenesis of insulin-dependent diabetes (53). Other studies confirmed and extended these findings (54).

Juvenile diabetes: a virally mediated disease? Mumps was first proposed as a cause of diabetes by Stang in 1864 (55). In 1926, a seasonal onset was observed for “acute diabetes,” fewer cases being diagnosed in May and June than in the spring and fall (56). In the following year, Gundersen reviewed mortality from early-onset diabetes in Norway before the introduction of insulin, noted that peaks coincided with outbreaks of mumps, and—in the knowledge that mumps could also cause pancreatitis—proposed mumps virus as the cause of acute diabetes (55). Intermittent reports continued to link mumps to diabetes, but interest in a viral etiology effectively lapsed until 1968, when the encephalomyocarditis virus was shown to induce transient diabetes in some adult mice but not in others (57). Soon after this, Gamble and colleagues published papers confirming the seasonal incidence of early-onset diabetes (58) and demonstrating an excess of neutralizing antibodies to Coxsackie virus B4 in patients with insulin-requiring diabetes who were tested within 3 months of disease onset and a subsequent fall in titers of these antibodies in the period after diagnosis (59). Both encephalomyocarditis and Coxsackie are picornaviruses. Coxsackie virus had long been known to produce morphological changes in mouse islets (60), and Gamble and Taylor went on to demonstrate that it could also induce diabetes in mice. Development of diabetes was strain-, age-, and sex-restricted, and the typical latency of 17–21 days between exposure to virus and onset of diabetes suggested a host response to viral infection rather than direct viral destruction of β-cells (61). It is characteristic of the elusive nature of the association between viral infection and diabetes that a negative study with the same viral strain followed quick on the heels of this report (62). At this stage, however, viral infection was still generally believed to lead to diabetes by cell lysis, and the concept that the effects of viruses might be both delayed and mediated by the immune system emerged more slowly (63,64).

Genetic heterogeneity

We have seen that the clinical heterogeneity of diabetes had been recognized for more than a century. The genetic heterogeneity of the disease was not to be established until the 1970s, at a time when the major players were more interested in the complications of diabetes than in its etiology. Juvenile and maturity-onset diabetes were conventionally regarded as “gradations or variants of the same basic disease with quantitative differences in insulin deficiency and hyperglycemia.” This view appeared to find strong support in the observation that the same complications—“accelerated atherosclerosis, microangiopathy, neuropathy, the complications of pregnancy, large babies and cataracts”—developed in both forms of the disease (65). The established paradigm maintained that the vascular complications of diabetes were determined by the gene or genes causing diabetes, and this view was maintained in the face of mounting evidence that metabolic factors were important. Those involved in this debate were therefore interested in defining genetic variants of diabetes and examining whether these all showed a similar pattern of late complications (supporting the metabolic hypothesis) or a heterogeneous pattern (consistent with the genetic hypothesis).

An impressive demonstration of genetic heterogeneity was to emerge from the study of twins. A previous study (66) had demonstrated that concordance approached 100% in monozygotic twins when the proband was diagnosed after the age of 40 years, as opposed to 25–50% when the proband was diagnosed before that age. In 1972, Tattersall and Pyke (67) confirmed these results in 96 monozygotic pairs, showing that 31 of 59 pairs diagnosed before the age of 40 years were concordant, as opposed to 34 of 37 after that age. More than 10 years had passed since diagnosis of the proband in the majority of discordant pairs, implying lasting protection from diabetes in the unaffected twin. This indicated that non-genetic factors must be of importance in the causation of juvenile diabetes and ran counter to the prevailing view that genetic factors were more important in early-onset than in late-onset diabetes. All this of course fitted well with the emerging hypothesis that early-onset diabetes was virally mediated. In retrospect, it can be seen that the timing of this often-cited paper and the emphatically stated conclusions were as important as the findings themselves, which were essentially similar to those of Gottlieb and Root (68) 4 years previously. These authors had already reported that 5 of 20 monozygotic twin pairs in which the proband developed diabetes before the age of 40 years were concordant for diabetes, as opposed to 10 of 10 monozygotic pairs with onset after the age of 40 years, and they had suggested that “it is in this older age group that a clear genetic contribution to the etiology is noted” (66).

Against this background, interest in the genetic heterogeneity of diabetes coincided with the emergence of the concept that juvenile diabetes had an autoimmune etiology. The HLA associations of early-onset diabetes and the identification of ICAs were therefore readily adopted as further support for the concept of heterogeneity (65,68), as was the description of a distinct genetic subtype known as maturity-onset diabetes of the young, which was not associated with HLA (69). The battle to establish that complications had a metabolic rather than a genetic cause was fought out in the “basement membrane controversy,” which found its most famous expression in a trio of editorials in the New England Journal of Medicine in 1976–1977 (70–72). Thus, and by an odd coincidence, 1976 saw editorials signaling acceptance of two fundamental concepts that have ruled our thinking ever since: the autoimmune nature of type 1 diabetes (14) and the metabolic basis for late complications (70).

Genetic susceptibility to diabetes. Meanwhile, another major paradigm shift had occurred in genetics. The mouse major histocompatibility complex (MHC) and the human HLA system were first identified as “transplantation antigens” by immunologists working toward human transplantation. Early speculation as to their biological function was given direction by the observation that there was genetic linkage between the mouse H-2 antigen complex and the susceptibility of the mouse to virus-induced leukemia (73). Immune response genes were identified and shown to be linked to the MHC in several animal species. By 1972, these loci had been shown to control the formation of specific immune responses to oncogenic viruses and exogenous antigens (74), and they were soon to provide the explanation for host specificity of response to viral infection. In 1973, Oldstone
et al. (75) showed that mice with different MHC types showed differing patterns of response to intracerebral injection of lymphocytic choriomeningitis virus, and soon after—but quite coincidentally—Zinkernagel and colleagues (76,77) demonstrated in the same model that T-cell recognition was MHC restricted. The association between autoimmune thyroiditis, susceptibility to lymphocytic choriomeningitis virus, and H-2 in the mouse, noted at around the same time (78), showed that immune-response genes could provide the link between susceptibility to viral infection and to autoimmunity, thus suggesting an important role in disease pathogenesis.

The search was on for HLA disease associations in man. An association with Hodgkin’s disease was noted in 1967, and by 1974, associations had been reported with systemic lupus erythmatous, ankylosing spondylitis, autoimmune liver disease, Graves’ disease, celiac disease, and multiple sclerosis, among other conditions (74). The first—and unsuccessful—search for HLA associations with juvenile diabetes was reported in 1972 by investigators who appreciated that diseases “most likely to reveal HLA correlations may be those showing features suggestive of a complex multifactorial genetic background, of a viral aetiology, or of autoimmunity” (79). These investigators examined 44 cases of juvenile-onset diabetes from an ethnically diverse population; ironically, they found (but did not comment on) an excess of HLA-A8 present in 10 of 28 of their Caucasian patients (79). In the following year, Singal and Blajchmann (80) published an association between juvenile diabetes and HLA-W15 (since known as B15)—although the investigators must have experienced typing difficulties because three alleles were demonstrated in some patients. Meanwhile Nerup had spent 18 months in a fruitless quest to demonstrate that diabetic patients had cellular hypersensitivity directed against viral antigens. The realization that genetic heterogeneity might explain the inconsistency of reports concerning viruses and diabetes prompted him to turn his attention to HLA genetics (J. Nerup, personal communication). In 1974, he published associations with HLA-B8 and -B15 (81). Cudworth and Woodrow (82) published the same observations in a letter very soon after, and they later reported an inverse relationship between age at diagnosis and HLA-A8, together with preferential zygotic assortment in sib pairs with diabetes, such that 10 of 15 were HLA identical versus the expected ratio of 0.25 (83). The concept that genes transmit the tendency to develop diabetes, but not the disease itself, was now established.

The discovery of ICAs. Identification of circulating autoantibodies preceded the demonstration of cell-mediated immunity in most autoimmune diseases, but juvenile diabetes was the exception. This was not for want of trying, and we have seen that many investigators had done so before 1974 (84). Their lack of success was partly due to poor illumination provided by earlier generations of microscopes. The introduction of new microscopes with epi-illumination made ICAs much easier to see. Islet immunofluorescence also escaped observation because it is weak relative to the bright staining seen with thyroid, gastric, and adrenal antibodies (85).

Serendipity played its role in the final discovery. Richard Lendrum, a research fellow in gastroenterology, wanted to look for antibodies to the exocrine pancreas in patients with acute and chronic pancreatitis, and came to Deborah Doniach for help and advice. Using frozen sections from good quality pancreas, in some cases obtained from cadaveric renal transplant donors, he observed that some sera produced a granular pattern of staining in sections from donors of blood group A and B, but not from group O, and that this was largely abolished by prior absorption with red cells of the same groups (R. Lendrum, personal communication). Group O pancreas was therefore best for immunofluorescent studies of the pancreas, and Lendrum was able to provide some freshly collected blocks to Franco Bottazzo, a research fellow newly arrived in Doniach’s laboratory. Bottazzo, who was finishing his thesis work on Addison’s disease, added pancreas to the panel of tissues already tested and found to his wonder that some of the islets lit up on the microscope stage (G.F. Bottazzo, personal communication). Bottazzo and Doniach went on to study 171 patients selected for the presence of organ-specific adrenal or thyroid autoimmunity with or without diabetes. Doniach’s interest in “polyendocrine” patients, in whom diabetes is associated with other autoimmune diseases, proved to be yet another happy coincidence, because ICAs in these patients stain very strongly and persist at high titer over many years. Strong staining reactions, present with rat and guinea pig as well as human islets, were seen in 13 patients, of whom 8 had diabetes at the time the sample was taken. Two of the remainder subsequently developed diabetes (84). Weeks later, MacCuish et al. (85) reported a study of 65 patients with juvenile-onset diabetes, 40 of whom were selected for other autoimmune disease or antibodies. ICAs were found in five, all of whom had Addison’s disease; four also had autoimmune thyroid disease. Both groups concluded from this experience that ICAs were rarely present in diabetes and then mainly in those with other autoimmune diseases. The characteristic β-cell–specific staining pattern and persistence of ICAs in these patients has since been explained by the presence of high levels of antibodies to GAD, which contribute to the pattern of islet immunofluorescence (86).

It soon emerged that islet autoantibodies are not uncommon. Using insulinoma cells as a substrate, Maclaren and Huang (87) detected circulating islet cell surface antibodies by indirect immunofluorescence in 34 of 39 insulin-dependent patients and speculated as to the possibility of immune intervention early in the course of the disease. Lendrum found ICAs in 85% of new-onset insulin-dependent patients and demonstrated that the prevalence of ICAs fell with increasing duration of disease. ICAs were also found in 5.3% of non-insulin-dependent patients and 1.7% of control subjects (88). Irvine drew attention to the presence of ICAs in patients on oral agents, in whom he found an 8% prevalence, and noted that these “had a strong tendency to develop IDDM in subsequent years, suggesting different grades of disease severity in autoimmune diabetes” (89). Zimmet et al. (90) made similar observations regarding the significance of GAD autoantibodies in late-onset diabetes 17 years later, leading to widespread interest in the condition then termed latent autoimmune diabetes of adults.
ANIMAL MODELS OF DIABETES
To some contemporary observers, characterization of juvenile diabetes as an autoimmune disorder remained incomplete until an animal model of immune-mediated diabetes could be developed. Two developments in 1976 remedied this deficiency. Streptozotocin was known to induce rapid necrosis of pancreatic β-cells when injected as a single high dose. Like and Rossini (91) developed a variant of this model in which multiple subdiabetogenic doses were administered. This was found to provoke insulitis in mice, associated with increased replication of viral particles within β-cells. Progressive insulin deficiency was noted well after streptozotocin had been cleared from the circulation, suggesting an immune-mediated phenomenon, whether mediated by virus activation or toxic damage to β-cells (91). In the same year, the Wistar-derived rat from the Bio Breeding Laboratories of Canada—the BB rat—was first described. These nonobese animals developed a spontaneous insulin-deficient, ketosis-prone form of diabetes associated with histological features of insulitis (92). The most widely used model of type 1 diabetes, the nonobese diabetic (NOD) mouse, was not described until 1980 (93) and did not become available to the international scientific community until some years later.

Reclassification of the diabetes syndrome. Before 1976, classification of diabetes was phenomenological rather than etiological. Some clinicians had attempted to distinguish stages in the natural history of diabetes (seen as a unitary condition) based on the degree of actual or potential glucose intolerance, and the American Diabetes Association proposed a four-step classification ranging from “prediabetes” through “subclinical diabetes,” “latent diabetes,” and “overt diabetes” (94). These categories are highly confusing to the modern reader and seem to have had the same effect on the physicians of the day. The terms juvenile and maturity-onset diabetes were more widely used, although as Irvine was to remark in 1977, “it is difficult to think of any condition in which the age of onset…or the type of treatment required are a satisfactory basis for classification” (63). These terms were later replaced by the more familiar IDDM and NIDDM.

Cudworth took the crucial step toward an etiological classification when he reintroduced the terms type 1 and type 2 diabetes in 1976 (14). He pointed out that type 1, although mainly a disease of early life, “can also occur as the insulin-dependent syndrome of typically abrupt onset at all ages, including the elderly.” Bottazzo et al. (95) rapidly adopted the new terminology and went on to propose (in an unconscious echo of the groups IA and IB of Dupurtuis) that type I diabetes (use of the roman numeral being most common at the time) could be further subdivided into type IA—caused by viral infection and associated with transient ICAs—and a polyendocrine form associated with persistent ICAs and to be known as type IB. Viral infection was still thought to account for most cases of diabetes. Irvine (63) next attempted to bring the two etiologies together by subclassification into type Ia (pure autoimmune) and type Ic (due to viral or toxic damage)—both considered rare—with Ib, the more typical form (due to both viral infection and autoimmune), as an intermediate between these two extremes. The current American Diabetes Association classification (96) of type 1 diabetes into IA (immune mediated) and IB (non–immune-mediated) bears some resemblance to Irvine’s suggestion.

The new paradigm. This retrospective presentation might make it appear that the development of these new concepts was a straightforward, logical, and even inevitable process. This is not how things would have seemed at the time. Imagine, for example, a modern grant review committee manned by conscientious and well-informed senior investigators and reviewing new proposals for 1973 in the light of what was known at the time. Nerup’s application to examine HLA associations in diabetes would stand little chance. A good group has just established that there are none (79), and Nerup is a young investigator with no experience in HLA and little track record in diabetes. As for Bottazzo, the best groups have looked for ICAs repeatedly and without success, and here is another inexperienced young scientist with no relevant track record. So it would go. As Kuhn pointed out, work developed from within a defined research area hardly ever results in a paradigm shift. Instead, the process is driven, as in the discovery of type 1 diabetes, by the incursion of techniques and ideas from other disciplines, headed up by young and untested investigators who bring no preconceived ideas to the subject.

When the new ideas came, they came in a rush, just as Kuhn would have predicted. Nerup’s HLA paper (81) was eventually published on 12 October 1974; Cudworth’s letter confirming these results (82) was published on 9 November. Bottazzo’s ICA paper (84) came out on 30 November, and MacCuish’s confirmation of ICA (85) came out on 28 December, accompanied by an editorial entitled “Autoimmune diabetes mellitus” (97). MacLaren published in May 1975 (87), soon followed by papers from Lendrum (88) and Irvine (89) establishing the basis of our current understanding of ICAs. All these papers were published in the Lancet, which therefore played a key role in promoting the new paradigm. The editor sat on Nerup’s paper for 9 months, finally yielding to a telephone call, but published MacLaren’s paper within 3 weeks of submission.

Although key individuals made key observations, the paradigm shift that resulted can be seen in retrospect as the culmination of a long, slow sea change in scientific opinion. Contributions by countless individuals in apparently unrelated fields had created the context in which the concept of autoimmune diabetes made immediate sense. Therefore, the climate of opinion had itself moved, in what might be termed a paradigm drift, creating the conditions that made the paradigm shift possible. Even so, not all became converted overnight. There is some truth in the saying of Max Planck that “a new scientific truth does not triumph by convincing its opponents […] but rather because its opponents eventually die, and a new generation grows up that is familiar with it.” The most important outcome was to galvanize a new generation of scientific investigators into action. In contrast, the impact of the new thinking on clinicians was muted, not least because they were more involved in the ongoing debate over control and complications.

The future of a paradigm. What is the current status of the paradigm? Knowledge about the disease process has accumulated at an explosive rate, and successive paradigm shifts have transformed thinking about the underlying molecular mechanisms. Techniques used by geneticists, immunologists, and virologists have converged to the point at which they are practicing variants of the same discipline. Paradoxically, and although there is no comparison between the state of knowledge 25 years ago and now, the disease paradigm...
DISCOVERY OF TYPE 1 DIABETES

has changed little over that time. Consider its first clear statement by Nerup et al. (81) in 1974:

One or more immune-response genes associated with HLA-A8 and/or W15 might be responsible for an altered T-lymphocyte response. The genetically determined host response could fail to eliminate an infecting virus (Coxsackie B4 and others) which in turn might destroy the pancreatic beta-cells or trigger an autoimmune reaction against the infected organ.

Given that 26 years have passed, that a revolution in molecular biology has taken place, and that some 15,000 papers relating to the etiology of type 1 diabetes have subsequently been published, this bears a remarkable resemblance to current statements of the paradigm.

The obvious explanation for the survival of this disease model, broadly similar to those proposed for other diseases in the human immune-mediated organ-specific family, is that it is correct. If so, the lack of detailed confirmation of the model, studied in so many diseases over so long a period, is remarkable. There seems no reasonable doubt that type 1 diabetes is immune mediated, but despite the strong association with the MHC, no necessary immune-response gene has been identified to date, a viral etiology has not been confirmed beyond doubt (98), and the concept of molecular mimicry remains tantalizingly elusive. Indeed, a recent review concluded that “no data convincingly demonstrate that mimicry is an important mechanism in the development of autoimmune disease in humans” (99). These observations do not invalidate the model, but may cause us to regard it in a new light.

One possible explanation, eloquently expressed in 1969 by Macfarlane Burnet (100) and reflecting the somber mood of his later years, is that autoimmunity is intrinsically heterogeneous. As he put it:

In all forms of autoimmune disease we find a basically similar situation—the appearance of a wide range of immunocytes reactive against the antigentic determinants involved and a complex and variable ecological situation dependent on the local availability of the antigens concerned and any mutant characters influencing the immunocytes’s response...the actual facts of autoimmune disease are so individual and only broadly reproducible that any approach must of necessity be a flexible one.

If this view, which might be called the heterogeneity model, is correct, the appropriate way forward might be to study type 1 diabetes in closely defined genetic and environmental contexts (for example, in Finland and Sardinia). Support for this view came from a recent genetic analysis of 1,551 families, which concluded that “our results illustrate the difficulties of genetic analysis of common disease...future studies should be carried out in large data sets from ethnically homogeneous populations” (101).

A more plausible variant might be a stochastic model. In this view, the number of variables involved is more limited, but the interaction between them is more complex. The outcome in terms of disease progression will therefore depend to a large extent on chance (probability), and the role of any single player in disease etiology will only make sense when considered in light of the movements of the other members of the team. If this view is correct, the way forward may lie in new forms of analysis (102). Finally, since the history of medicine teaches us that simple solutions have often been found to seemingly complex and intractable problems, the quest to define simple unifying explanations must and will continue.

Kuhn emphasized that paradigms do not age gracefully. Instead they tend to ossify, to hinder thought instead of freeing it. Inconsistencies tend to be ignored or explained away but nonetheless accumulate until the established model begins to show signs of strain. Meanwhile, the process he refers to as “normal science” is typically paradigm driven. Although he did not put it this way, in normal science the emphasis is on extending familiar solutions to new problems rather than the other way around. This may be why Richard Doll compared the current belief that massive research funding can solve fundamental scientific questions rapidly to the notion that a baby can be produced within 1 month by the expedient of making nine women pregnant (103).

The ultimate aim of research in type 1 diabetes—at least from the point of view of a clinician—is to explain the human disease and to show us how to prevent and treat it effectively. If this excursion into the process of discovery teaches us anything, it is to cherish inconsistencies as potential growth points for the future. It makes sense to focus on areas where the conventional paradigm fails to explain the human disease, rather than where it succeeds. Why is the incidence of diabetes in children rising so rapidly (104)? Why does the age of the mother at delivery appear to impact on the risk of diabetes in the child (105)? The discovery of type 1 diabetes is an ongoing process, not an historical event, and the disease itself will surely show us the right way to go.

ACKNOWLEDGMENTS

Although he will not bother to read it, this article is dedicated to the colleague who, when informed of this project, remarked that he was “not yet old enough to take up an interest in the history of science.” Many people have discussed this project with me or offered comments on the paper, and particular thanks are due to Jean-Francois Bach, Franco Botazzo, Richard Lendrum, John Lister, Ian Mackay, Noel Maclaren, Jorn Nerup, and Robert Tattersall. Errors of fact, attribution, or interpretation are my own, and I would welcome correspondence on such issues. The phrase “the Geneticist’s nightmare” was coined by J.V. Neel.

REFERENCES


