How did the study come about?
Schizophrenia is a severely disabling disorder with a lifetime risk of 1% and a point prevalence of about 5 per 1000.\(^1\) Schizophrenia accounts for 2.3% of the global burden of disease.\(^2\) The aetiology of schizophrenia is complex. While there appears to be a significant genetic contribution,\(^3,4\) it is possible that many genes of small effect contribute to the disorder and that environmental risk factors interact with the genetic susceptibility. The impact of environmental risks appears to be complex, with no single environmental risk factor of major effect isolated to date. Putative candidate exposures include winter/spring season of birth, pregnancy and birth complications; prenatal starvation; urbanicity; migration and ethnicity, with the incidence apparently higher in second-generation migrants; stress and cannabis use. There have also been reports of possible age-related germ-line mutations such as those associated with older fathers.\(^5\) Much of the research into environmental risk factors has focused on the role of obstetric complications. While the level of reported risk factors has focused on the role of obstetric complications. While the level of reported risk factors has focused on the role of obstetric complications varies, Geddes and Lawrie, in a meta-analysis of published research, reported a pooled odds ratio of 2,\(^6\) a figure corroborated a few years later by Cannon et al., who calculated the individual odds ratio of 2 and under for the majority of the many complications that they reviewed.\(^7\) It appears that changes in the uterine environment such as increased maternal cytokines in response to hypoxia\(^8\) or infection,\(^9\) and stress-induced elevation of circulating corticosteroid levels\(^10\) may predispose some infants to psychopathology. In particular, much attention has been focused on hypoxic/ischemic events associated with and/or resulting in placental abnormalities leading to fetal growth retardation and neonatal encephalopathy. In order to examine the role of genes, environment and their interaction in the expression of schizophrenia, a number of studies have adopted a prospective, longitudinal approach focusing on so-called ‘strategic’ populations\(^11\) of ‘high-risk’ children. These are children who are at genetically increased risk of developing schizophrenia because they have at least one parent with the disease and are therefore a priori enriched for predisposing genes. High-risk samples offer researchers the opportunity to address efficiently and parsimoniously key questions about risk factors and vulnerability. By parcelling out the separate contribution of genetic liability and environmental exposures, their relative contribution to the onset of schizophrenia and other neuropsychiatric outcomes can be assessed. One of the earliest of the longitudinal investigations of developmental pathways in the children of parents with schizophrenia was initiated by Barbara Fish in the early 1950s. Fish proposed that developmental retardation and neurological soft signs observed amongst the high-risk children in her sample were early markers of an inherited neurointegrative defect in schizophrenia which she named ‘pandysmaturation’.\(^12\) Dr Fish’s seminal work on pandysmaturation laid the groundwork for the formulation of the neurodevelopmental hypothesis of schizophrenia in the late 1980s. In its simplest form, the hypothesis states that critical circuits in the brain are affected by subtle disease processes in early development, with full-blown consequences
becoming evident as schizophrenia when the affected brain areas reach physiological maturity in adolescence or early adulthood. The original theory was open as to whether the disease process was genetically or environmentally determined. This theory continues to exert a powerful influence on our understanding of the aetiological basis of schizophrenia. Subsequently, other prospectively assessed cohorts were established. These include, among others, the Copenhagen High Risk Project in 1962, the Israeli High-Risk Study, the Swedish High-Risk Study, and the New York High Risk Project, many of which have been followed up for a very long time. Comprehensive reviews of the key high-risk studies up to the end of the 1990s are available. Although these are highly regarded studies, they are subject to one or more important limitations, including reliance on selected samples of small to modest size, restriction to the children of mothers with schizophrenia only and absence of prospectively collected data on early environmental impacts due to cohort formation in middle childhood or later.

Population databases with access to networks of longstanding, whole-of-population administrative registers including psychiatric case registers, offer an alternative approach to conventional high-risk studies. In such jurisdictions, it may be possible to establish retrospectively an electronic cohort of high-risk children using prospectively collected data across linked registers. Children born to unaffected mothers form the comparison. For example, Bennedsen and colleagues examined the risk of obstetric complications, congenital malformations and mortality in the children of mothers with schizophrenia using the Danish registers. Lichtenstein and colleagues linked the Swedish multigeneration register for the individuals born 1932–2002 to the hospital discharge register to examine genetic and environmental contributions to liability for schizophrenia and bipolar disorder. We have taken a similar approach in Western Australia, where we have set up a longitudinal, multigenerational high-risk e-Cohort to answer questions on the relative contribution of familial liability and environmental exposures at different developmental stages in the aetiology of schizophrenia and other neuropsychiatric outcomes. While a key focus is on the role of obstetric complications, we include an extensive range of covariates and potential mediators along the developmental pathway.

What does it cover?

The e-Cohort has been constructed by record linkage across Western Australian administrative health and social services registers. The development of Western Australia’s record linkage capacity for health research is described elsewhere. Western Australia offers several advantages for register research. It is physically isolated from other population centres by natural boundaries: the Nullarbor Plain and desert to the east and north, and ocean to the west and south. Western Australia’s strong economic position has resulted in relatively low levels of out-migration and high levels of in-migration. It is a highly urbanised State: in 2008, 73% (1.6 million) of its population of 2.2 million was living within the greater metropolitan area of its capital city, Perth.

The high-risk e-Cohort is composed of all the children born in Western Australia between 1980 and 2001 to mothers with a diagnosis of psychosis (schizophrenia, unipolar major depression, bipolar disorder and other non-organic psychoses) who have had an inpatient or outpatient contact recorded on the statewide psychiatric case register between 1966 and 2001. The inclusion of mothers with psychoses other than schizophrenia allows the researcher to investigate the specificity of findings across the psychosis spectrum. The comparison cohort consists of all the children born in Western Australia over the same period.

Figure 1 Model of the impact of genetic and environmental factors on the neuropsychiatric phenotype at various developmental stages
period whose mothers have not had a record of contact with mental health services. Fathers have been identified and sibships established through probabilistic record linkage of birth records. Access to the Western Australian linked registers has enabled the research team to develop complex profiles of parental psychiatric and physical morbidity. Detailed data on the pregnancy and perinatal period have been extracted from the Midwives Notification System. Additional information on all children has been compiled from multiple registers. The registers cover all hospital admissions, inpatient and outpatient psychiatric contacts, intellectual disability, birth defects, cerebral palsy, cancer and mortality. Other potential data sources have been identified in the study protocol and are being added as they become available. Most recently, linkages have been included on school assessment data, including literacy and numeracy proficiency, and data from state departments on contacts with child protection services and corrective services. The creation of the e-Cohort has been funded since 1996 by a series of Australian and US grants, including Stanley Foundation (USA), National Health and Medical Research Council (Australia) and March of Dimes (USA).

The study design and data collection protocol enable researchers to (i) identify the spectrum of morbidity, developmental outcomes and antecedents of psychopathology, as well as positive developmental pathways, in children at high genetic and environmental risk at critical time points from birth to young adulthood, (ii) examine the relationship between such outcomes and specific risk factors operating at different developmental stages, but particularly targeting genetic liability and obstetric complications and (iii) identify similarities and differences in offspring outcomes for mothers across several diagnostic groups, namely schizophrenia, bipolar disorder and unipolar major depression. The study also provides a unique evidence base for better informed interventions and management strategies, including risk reduction through targeted antenatal and postnatal interventions.

**Who is the sample and how often have they been followed up?**

The e-Cohort was conceived by Professor Assen Jablensky in the mid-1990s. The original cohort (Phase 1) consists of 6303 children born in Western Australia between 1980 and 1992, comparing 3174 children born to 1831 case mothers with psychosis (including 618 births to 328 mothers with schizophrenia) and 3129 children born to 1831 randomly sampled comparison mothers with no known contact with psychiatric services. In Phase 2, the e-Cohort was expanded to include all births from 1980 to 2001, covering 20 209 children born to 9750 case mothers and 45 251 children born to 239 366 comparison mothers (see Figure 2). Fathers of the children have been identified as part of a larger Western Australia project. Sibships among offspring have been constructed and determination of the family relationship (e.g. full sibling or half sibling) is possible. Some children have become parents themselves, so we are now collecting data for three generations. To date, we have 3391 children in the third generation born to 3150 parents in the second generation, with 6028 grandparents. These numbers continue to grow over time.

There are no specified follow-up time points. Rather, data collection is cumulative. Key data for an individual are prospectively collected on individual registers and may be retrospectively updated on the research database to include all contacts made and recorded on the linked registers.

**What has been measured?**

The key measures collected for the e-Cohort are the children’s high-risk status (i.e. having a mother with psychosis), their exposure to obstetric complications in utero or at birth, a range of additional exposures along the developmental pathway and the children’s neuropsychiatric outcomes.

**Psychiatric morbidity**

Data on psychiatric morbidity have been extracted from the psychiatric case register, which dates back to 1966 and, unlike many psychiatric registers, is not restricted to inpatient admissions but also includes outpatient and community care contacts. The register uses the International Classification of Diseases (ICD) to record diagnosis at multiple points of contact over the course of an episode of care. For the current programme of work, an assessment of psychiatric morbidity is made when a person has at least one ICD-8 or ICD-9 code in the Chapter 5 range or an ICD-10 code in the F range, signifying a diagnosis of mental illness. Specific disorders within the psychiatric spectrum are identifiable. To accommodate changes in diagnostic classification over time, all codes have been translated to the International Classification of Diseases, 9th Revision. The ICD-9 codes used for psychosis are schizophrenia (ICD-9 295.x), bipolar disorder (ICD-9 296.0 and 296.2–5), unipolar (major) depression (ICD-9 296.1, 296.6, 296.8 and 296.9) and any non-organic psychosis (ICD-9 295.x, 296.x, 297.x and 298.x). The research team has developed and validated several algorithms for determining a diagnosis of psychosis for an individual who has had a psychotic episode using the multiple records for that person on the psychiatric case register.
Other neuropsychiatric outcomes

Intellectual disability is determined using the Intellectual Disability Exploring Answers (IDEA) database. The fields on the database come from two sources. The main data source is the Western Australian Intellectual Disability Register, which includes all the individuals in the state who are registered for disability services. Since registration is required for receipt of services, coverage is good and increases with age. American Association for Mental Deficiency criteria are used to define intellectual disability, namely an IQ that is two or more standard deviations below the population mean and related limitations in adaptive behaviours and skills. These data are supplemented with records for school children who have been identified as having an intellectual disability and are therefore eligible for special education services. Level of intellectual disability is recorded for all individuals. Rare syndromes and birth defects are identifiable through both the IDEA database and the Birth Defects Registry. The latter covers malformations diagnosed up to the age of 6 years and includes live births, stillbirths and pregnancies less than 20 weeks’ gestation terminated because of fetal malformation. A more general assessment of cognitive capacity is available for the children born between 1989 and 1996 using statewide educational assessment data for literacy and numeracy. Approximately 87% of children on our database born in this period have had at least one assessment.

Obstetric complications

Core data for each birth are recorded in the Midwives Notification System, which includes all births in Western Australia of $\geq 400$ g or 20 weeks’ gestation since 1980. For the mother, the fields include, among others, age, marital status, race, country of birth, height, previous pregnancies and their outcomes, pregnancy details including the date of the last menstrual period and the expected due date. Specific complications recorded prospectively on the Midwives Notification System include pregnancy complications (e.g. threatened abortion, essential hypertension, preclampsia, placenta praevia, abruptio placenta, other antepartum haemorrhage, drug side effects), labour and delivery complications (e.g. threatened preterm labour, prolonged labour, precipitate delivery, cephalopelvic disproportion, malposition, prolapsed cord, postpartum haemorrhage) and early neonatal
complications (e.g. fetal distress, 5-min Apgar score, time to spontaneous respiration, intubation, use of a narcotic antagonist).

The McNeil–Sjöström Scale for obstetric complications has been used to provide composite scale scores capturing the timing, number and severity of complications experienced. The McNeil–Sjöström Scale was selected because its scoring of complications is underpinned by strong biological considerations, including the timing of the adverse event in pregnancy and its likely impact on the development of the central nervous system. Moreover, it allows the researcher the flexibility of setting a minimum threshold for the inclusion of obstetric complications. For example, most analyses that have been undertaken on the e-Cohort use a relatively high threshold of severity level 4. The severity levels range from 1 to 6, with complications at level 4 and higher covering complications that are ‘potentially clearly harmful or relevant’. Scores are provided for complications occurring in the pregnancy period, at the time of labour and delivery and in the neonatal period. Since 1996, the research team has worked closely with the developer of the scale, Professor Thomas McNeil, and has produced a validated computer algorithm based on what is, essentially, a manual scoring system. As a consequence, we now have the capacity to score efficiently and comprehensively the many complications stored in electronic format in half a million midwives records.

There have been a number of additional enhancements. These include (i) the inclusion of longitudinal measures of maternal morbidity, such as diabetes, asthma, essential hypertension, chronic renal impairment, epilepsy and thyroid dysfunction, so that relevant medical information recorded for a mother in one pregnancy is carried over to subsequent pregnancies and (ii) ongoing work on the construction of a validated measure of neonatal encephalopathy using register data.

Other exposures
In the current programme of work, a critical task is to construct measures of psychosocial adversity and other developmental exposures (such as separation, loss and stress) using register data. The constructs being developed include measures of, among others, disrupted family structure, separation from parents, serious illness in a parent or sibling and indicators of social disorganization within the family, such as residential mobility. The resultant constructs will be validated against clinical data. In addition, the database includes data on child abuse allegations from the Department of Child Protection. This includes the type of abuse (sexual, physical, emotional and neglect), the person responsible, the informant and the allegation outcome. The percentage of children with at least one allegation is 5.7%. Also available for analysis are measures of socio-economic status. In addition to individual-level data for socio-economic status, we use area-level data based on the address at the time of birth. The latter are described in greater detail below.

Additional measures from clinical case notes
A unique feature of the e-Cohort data collection is the supplementation of register data with data manually extracted from clinical case notes, coded and stored electronically on the study database. This complements the administrative data with a rich source of qualitative information. These data have been collected for a sample of mothers, for all children with at least one record registering a diagnosis of schizophrenia and for a sample of children with other diagnoses. Three additional datasets have been created as a result of the intensive clinical case-note reviews. These are the Diagnostic Interview for Psychosis database that records psychopathology and assigns provisional diagnoses; the Children’s Checklist data collection designed specifically for the e-Cohort to record data on substance misuse, prescribed medication use, behavioural problems, neurocognitive problems and other psychopathology not covered in the Diagnostic Interview for Psychosis and the Mothers’ Medications database to describe prescribed medication and illicit substance/alcohol use 6 months before conception and during pregnancy. Currently, the data from the children’s case notes are being used to map, graphically, events of clinical significance along a child’s developmental trajectory. A similar approach is applied to the information from the register data. These trajectories are then overlaid onto timelines—see Figure 3. Used in this way, the charts are a powerful, visual tool, reconstructing key events along the life course and mapping their interdigitation.

What has it found? Key findings and publications
Publications to date have been based on Phase 1 data for children born between 1980 and 1992. Key findings indicate that mothers with schizophrenia, bipolar disorder and unipolar depression experience an increased overall incidence of obstetric complications, relative to the non-psychiatric comparison group. Excess of obstetric complications is observed mainly in births occurring post-onset of maternal psychosis, suggesting a role for behavioural disorganization and environmental exposures such as poor nutrition and substance use. Outcomes significantly increased in women with schizophrenia but not women with bipolar disorder or unipolar depression, include placental abnormalities, low birth weight, minor physical anomalies, cardiovascular birth defects and offspring diabetes. The rates of these outcomes show no pre-onset/post-onset differences, suggesting a
diagnosis-specific pre-existing susceptibility that may involve both genetic and environmental components. Work in progress indicates that rates of neuropsychiatric outcomes (rare syndromes, intellectual disability, psychiatric morbidity including psychosis) in high-risk children are well above population rates. Furthermore, although both familial liability for neuropsychiatric disorders (intellectual disability; psychiatric morbidity; psychosis) and exposure to obstetric complications are significantly associated with the range of outcomes, they appear to act independently.

The study findings inform clinical and public health practice, as well as aetiological research. An important outcome to date has been the translation of these research findings into clinical practice through a Western Australian Department of Health grant to design and evaluate an antenatal care intervention programme for women with severe mental illness. As a result, in 2008, a specialist Childbirth and Mental Illness antenatal clinic, the first of its kind in Australia, was established at the State’s main maternity hospital with the intervention package implemented in community mental health services throughout the state.

What are the strengths and weaknesses of the e-Cohort data?

This register-based e-Cohort offers the researcher some important advantages over other prospective or retrospective data collections. It is a whole-of-population e-Cohort, utilizing prospectively collected data from longstanding administrative registers that have not been affected by periods of discontinuity. The identification of genealogies in the linked data enables research into genetic influences, and gene–environment interactions. It also permits the use of sibling comparisons in addition to unaffected population comparisons. The size of the e-Cohort ensures sufficient power for most statistical purposes; this is of particular importance when examining rare outcomes. In addition, the use of longitudinal data collected over a long period means that outcomes distal from the exposures of interest are included. Moreover, this methodology avoids reliance on a person’s recall of events and the consequent risk of recall bias. The inclusion of data extracted from clinical case notes adds a dimension not usually found in record linkage studies.

There are, however, several biases and omissions that may affect these data. One of these is loss to follow-up. Loss to follow-up due to death in Western Australia is not a problem since the mortality register is routinely linked to other Western Australian registers. A more serious loss to follow-up may arise as a result of movement out of the state and, hence, beyond the geographic ambit of the registers. This is especially problematic if loss through out-migration affects some groups within a population more than others. In Western Australia, the State’s natural advantages in addition to a relatively thriving economy has given it one of the highest rates of Australian interstate in-migration in the past years and, combined with its geographic isolation, one of the lowest levels of out-migration. It has been
estimated that, in 1994, permanent out-migration was 1.7% per year compared with an aggregate of 2.1% for the other states and territories combined.\textsuperscript{28} Overall, out-migration of any kind has remained fairly constant, averaging 2.7% for the period 1990–2008, whereas in-migration has increased slightly, averaging 3.6% over the same period. None of the available evidence suggests that out-migration introduces a systematic bias into the data.

A further source of bias may arise if there are ‘failure to link’ errors across the registers. This may be problematic in a country such as Australia that does not have a national system of unique identifiers for its population. A probabilistic data–matching methodology is used to link records across Western Australian health registers.\textsuperscript{28} The accuracy of record linkage using probabilistic data matching is high, with the percentage of true matches estimated to be 97–99\% depending on the number of matching variables used.\textsuperscript{37} Another possible bias arises if there are specific groups of people on the registers who are prone to ‘failure to link’ errors. For example, records for women may be missed because of surname changes at the time of marriage or the breakdown of a relationship, and especially for those who do not have Anglo-Saxon surnames, through misspelling of the name. Another source of linkage failure is the use of aliases, more commonly by persons in contact with mental health services. Linkage failure is minimized in these cases through the recording of maiden names on the Midwives Notification System, the keeping of files of aliases on the psychiatric case register and the use of sound compression algorithms at the time of linking to pick up variant spellings of names.

A potential drawback in the use of a register-based e-cohort is the restriction to data on registers that have been established primarily to collect information for administrative, not research, purposes. Consequently, confounders and modifiers such as socio-economic status, family environment and other childhood exposures may be missing. With the Western Australian Schizophrenia High Risk e-Cohort, we have been able to overcome at least some of these limitations. First, socio-demographic data are available on several electronic registers; this includes data on sex and race. Access to multiple sources for these two variables not only enables cross validation of data but also facilitates the assignment of values where data are missing on one register but not on another. Data on occupation and education, but not income, are also available on a number of the registers, but the quality of these data is variable and data are not available for every one. Consequently, particular use has been made of the Socio-Economic Indexes for Areas (SEIFA), measures constructed by the Australian Bureau of Statistics using principal components analysis of census data for more than 40 questions of social and economic interest.\textsuperscript{38} The basic unit of analysis of a SEIFA is a spatial unit known as the census collector’s district, a small geographic area consisting of 250 dwellings on average. Where possible, the spatial unit used to assign SEIFA is the small census collector’s district to avoid misclassification when using area-based characteristics.\textsuperscript{38–40} If this is not available, postcode, which covers a larger spatial area, is used. SEIFA data are available for 99.5\% of births on the study database. Second, our registers do not inform us about factors such as childhood adversity or conditions such as neonatal encephalopathy. Consequently, it has been a priority for the research team to operationalize difficult constructs and validate the resultant proxies against clinical data. This work is currently underway and will greatly enhance the research database. Third, lifestyle variables are either not available on the registers or, if available, are recorded intermittently and are generally of uncertain quality. In particular, antenatal data on maternal nutrition, smoking, substance abuse and exposure to psychotropic medication in pregnancy, all critical to disentangling the separate contributions of genetic and environmental factors to the outcomes, are not available for all births. Where possible, the statistical impact of these missing data can be modelled using estimates based on data available from alternative sources. This approach has been used to examine the potential impact of smoking in pregnancy by women with schizophrenia compared with comparison mothers.\textsuperscript{32} In addition, the access to clinical case notes provides a unique source of descriptive data rarely found on administrative registers, including medication data and their timing, albeit for a limited sample.

Can I get hold of the data? Where can I find out more?

Current access to the data is restricted to researchers named on the program’s ethics application. Potential postgraduate students are invited to contact the principal investigator (A.J.). It is anticipated that, once the data have been cleaned and key constructs fully operationalized, other researchers will be welcome to work with senior investigators on the study, subject to ethics approval.

Funding

Theodore and Vada Stanley Foundation (USA), National Health and Medical Research Council (Australia) and March of Dimes (USA).

Acknowledgements

The authors acknowledge the invaluable contribution of the skilled team of mental health research nurses...
including Carmel Ryan, Cate Simpson, Peterina Duggan and John Dean who extracted data from the clinical case notes, and that of Patricia Di Prinizio who is currently developing constructs based on register data. In particular, we note the significant scientific contribution of Professor Carol Bower and Professor Stephen Zubrick who have collaborated with us on this programme of work since its inception in the mid-1990s.

Conflict of interest: None declared.

References


