6

# Reproductive and Developmental

**Robert E. Chapin, Sally Perreault-Darney, and George P. Daston**

*ABSTRACT': Apical tests for reproductive or developmental toxicity assess the potential for a compound to affect any of the thousands of steps involved in making gametes and in the successful development of a fully functional offspring. Conventionally, this is thought to require at least 21 days for rodent female reproduction and for development, and close to 70 days for spermatogenesis. Short-term tests can evaluate some subset of these processes, so multiple tests must be used. The best use of in vitro tests currently is for evaluating a series of structurally-related molecules with an endpoint which reports a specific type of toxicity known to affect at least some members of that class. Because no in vitro tests have been found to correlate well with the breadth of reproductive and developmental toxicity observed in vivo, test-tube or culture-based tests should not be used as a first-pass, general screen for these effects. Even though short-term (21 or 28 day) in vivo studies will miss a variety of transgenerational effects, they remain the best means of identifying the more potent developmental and reproductive toxicants.*

We will review the rationale for the current versions of definitive tests for reproductive and developmental toxicity, the approaches taken in reducing the duration of these tests and documenting what is gained and lost by such alternatives. Finally, we will address in vitro and genotoxicity tests, and review briefly their advantages and shortcomings, and their relationship to in vivo developmental/reproductive toxicity results. To be explicit, this consideration moves from the best to the worst, in terms of confidence in the information generated.

It is the feeling of this group that good screens for toxicity evaluate as much of a process at once as possible. This is the standard against which we will judge the value of a potential screen.

## ▮ DEFINITIVE TESTS FOR REPRODUCTIVE TOXICITY

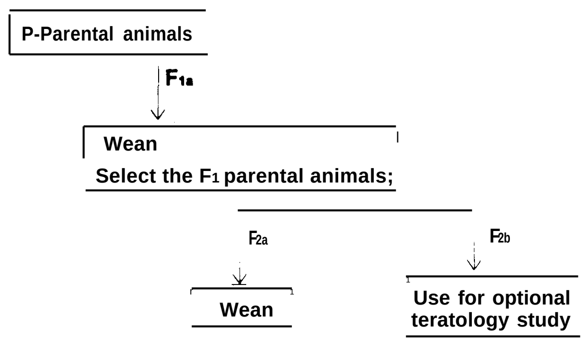
Definitive tests for reproduction and development are, essentially, set up to maximize confidence in a negative result. The definitive *in vivo* tests are apical, that is, they evaluate the integrated function of the entire system in one test. The benefit: if the results are negative, then one has reasonable assurance that there has been no effect anywhere in the process. The down-side is that identifying the location of a lesion or adverse effect can be slightly more time-consuming when starting from apical data.

As apical tests, these designs expose the entire process to the toxicant in question. Although this also depends on the pharmacokinetics of the compound in question, some default durations have been evolved, based on biology: For female rodent reproduction, the adult females should be exposed for 3-4 weeks prior to conception, as this exposes 4 or 5 estrous cycles of 5 or 4 days, respectively. In practice, the females may be exposed for 2-3 times this length of time, but 21-28 days is generally considered the minimum.

For spermatogenesis, the concept is to expose the gamete from a spermatogonium until it is ejaculated (again, the concept of exposing all stages of the process to the toxicant). In the rat, this is approximately 60-70 days. In practice, this often winds up being a 90 day exposure.

The field of developmental toxicology is in transition. In the past, the concept held that most terata resulted from exposures during organogenesis. In the rat, this begins about 6 days after mating, and continues until gdl5 (gestation day 15, out of a 22 day gestation period). Recent evidence shows that significant effects on the fetus can occur shortly after fertilization, and well before implantation (which occurs approximately on gd6), so many newer studies begin exposure of the pregnant female the day after mating, and continue until the day before delivery, when she

Figure 6-1: Sequence of a Two Generation Reproduction and Teratology Study


P-Parental animals
F1a
Wean
Select the F1 parental animals;
F2a
F2b
Wean
Use for optional teratology study

is killed and her fetuses examined for structural abnormalities. Alternatively (or additionally), some countries require testing of the offspring for behavioral abnormalities ("behavioral teratology"), which requires that in utero exposure be followed by 1-3 months of testing post-partum. Excepting these behavioral tests, this means a maximum of a 3 week exposure period for structural developmental toxicity.

Additionally, we should acknowledge that, while scientists have divided up the process into fields of specialization based on gender or process, Mother Nature is not so cut-and-dried. There is a considerable (and increasing) body of evidence that adverse reproductive effects on the offspring can be produced by a single in utero exposure to some compounds, or even by treating the adult parent before pregnancy is initiated. That is, treating a pregnant female with hormonally active compounds can produce permanent changes in her offspring. In the case of the reproductive system, these changes will not, of course, be visible until those offspring start to reproduce, a time lag of 2.5 months (rodents) to 16-20 years (humans). In these cases, the division between reproductive tox and developmental tox is well and truly blurred. (Note that this also can occur with systems other than reproduction; the post-natal manifestations of pre-natal exposure can be delayed until well into that offspring's life.)

This occurs because of the concept of "critical periods". Each organ system, as it develops, passes through a period (or multiple periods) where the signals received from (or through) the mother determine the long-term status of that system. This set-point is only adjustable for a short period (the "window" opens only briefly). The animal is vastly more sensitive to exposures while the window is open than at any other time in its life. For example, short exposures to TCDD at a specific point in gestation will permanently reduce the size of the gonads or the number of ovarian follicles. Slightly too much thyroid hormone (or a toxicant that mimics thyroid hormone) will have a similar effect, while too little thyroid hormone at a critical period will remove the signal to stop dividing, and testes in the adult will be permanently enlarged (perfectly normal, producing functional sperm, just bigger). This occurs for other systems as well: limited exposure to PCBs will permanently reduce the levels of circulating vitamin A in the kids, an effect with unknown consequences. The important concepts here are that: 1) the developing organism passes through some windows of vulnerability that do not exist in adults; and 2) changes made during these times can have permanent consequences for the offspring. The implications: 1) apical tests will (by definition) continue exposure during these times, and 2) short-term tests that ignore these windows increase the likelihood of missing a potentially significant toxic effect.

In practice, all of the previous considerations are folded together into a multigenerational test (figure 6-1) that starts off with either adult or pubertal animals (generally rats, and 20-30 of each sex per dose level), and exposes them to the toxicant in question for approximately 70-90 days, and then mates them within a treatment group (the high dose males mated to the high dose females, etc). Treatment continues while the dams are pregnant, after they have delivered, and then after weaning, the offspring are treated with the same dose their parents received. The pups are treated until they are about 70-80 days of age, when they are mated (again, within treatment levels), and another generation is produced. This second round of pups is killed either shortly after birth, or at weaning. In theory, this strategy should allow a compound to be identified as toxic no matter where in the reproductive process it works (Recognize that senescence of reproductive function is not being examined in this scenario, and it is probable that a compound that reduces the number of ovarian follicles will not show up functionally, because the reproductive lifespan of the animal is not being assessed, only the beginning of the process is tested. It is possible that counting ovarian follicles may identify premature follicle loss, but this is rarely done.) The test is apical: it evaluates the entire process of reproduction, from stem cell gamete through finished pup, to the reproductive capabilities of that pup as an adult. It identifies heritable damage (to the gamete's DNA), as well as effects on lactation, parturition, etc.

Variations on this theme are common: two litters can be produced per generation, and one can be reared to evaluate second-generational effects, while the second can be assessed for structural abnormalities. The National Toxicology Program's (NTP) Reproductive Assessment by Continuous Breeding protocol (RACB) is more of a forced-breeding design, generating 4-5 litters in the first generation. The idea is that if the system is "pushed", adverse effects are more likely to be identified. Additionally, the extra litters take no more time, and produce vast increases in the statistical power to identify toxic effects. The Alternative Reproductive Test, developed by the U.S. Environmental Protection Agency's (EPA) Health Effects Research Lab, starts dosing the first generation at weaning, and generates several litters from the second generation. This maximizes the exposure of juveniles to the toxicants, a time period when hormonally-sensitive windows are known to be open.

Inherent in all these protocols are cell- and tissue-based assessments of the reproductive system at necropsy. This is necessary because fertility can be normal even though there are measurable reductions in, for example, gamete number: sperm count must be reduced significantly (by 50°/0-900/0) to reduce fertility in a male, while fewer follicles in a female rodent will not show up as reduced fertility until 4-7 months of breeding. So, conjoint with the in vivo fertility assessments are specific evaluations of the systems at necropsy (sperm measures, ovarian follicle counts, histopathology, etc). This disassembles the system some, providing preliminary information on the site of effect.

It is also important to note that these necropsy endpoints (organ weights, sperm assessments, estrous cyclicity) can all be added to the end of a 90 day subchronic test. This strategy is used routinely by the NTP to identify probable reproductive toxicants, and those compounds that deserve more definitive testing for reproductive toxicity.

To summarize: definitive tests for reproductive toxicity strive to expose all parts of the reproductive process to the putative toxicant. If no adverse effects are seen, there is some confidence that human risk from exposure to such a compound will likely be low.

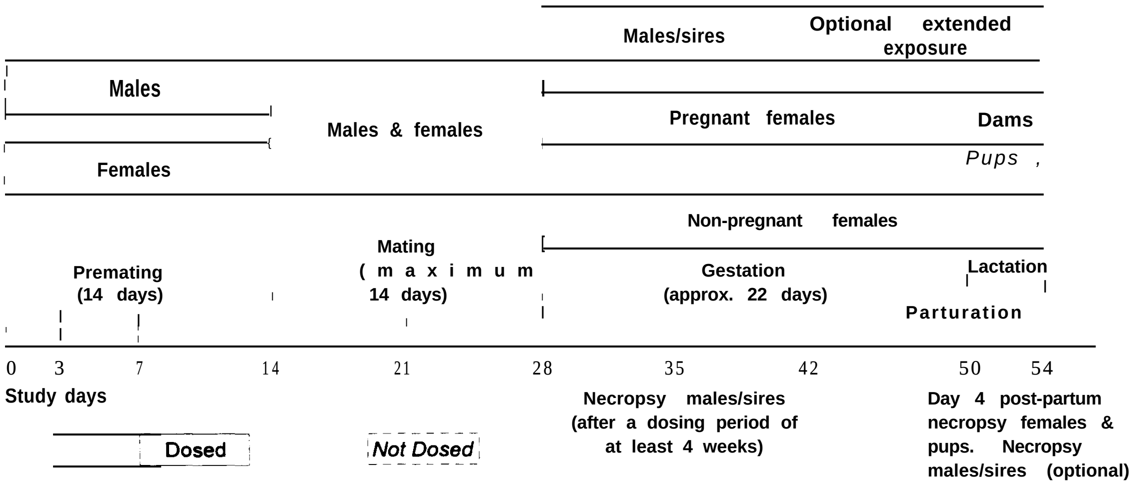
## SHORT TERM TESTS

The greatest gains in reducing the duration of testing come from reducing the duration of exposure of the male, since the female and developmental toxicity portions of the definitive test are only about 3 weeks long. Thus, the short-term tests described below tend to truncate the male portion of the process the most, and make less drastic changes in identifying effects in females or fetuses.

An additional theme will become evident below: if a single apical test is going to be replaced, those replacements must be multiple, wherein each examines an individual part of the system (be it female reproduction or development or male reproduction). That is, a group of shorter tests can be acceptable if each component of the process is evaluated individuallv.

There have been a few designs evaluated for short-term assessments. The Organization for Economic Cooperation and Development (OECD) has recently sponsored a workshop which generated a shortened test for developmental and reproductive toxicity (figure 6-2). In this test, pairs of rats are treated for the duration of the test, which is approximately 54 days. Fourteen days after the start of treatment, males and

Figure 6-2: Experimental Schedule Indicating the Maximum Study Duration, Based on 14-Day Mating Period


Males/sires
Optional extended exposure
Males
Males & females
Pregnant females
Dams
Females
Pups
Non-pregnant females
Mating (maximum 14 days)
Premating (14 days)
Gestation (approx. 22 days)
Lactation
Parturation
0
3
7
14
21
28
35
42
50
54
Study days
Dosed
Not Dosed
Necropsy males/sires (after a dosing period of at least 4 weeks)
Day 4 post-partum necropsy females & pups. Necropsy males/sires (optional)

females are cohabited for up to 2 weeks, and the females are allowed to gestate and deliver their young. The young are evaluated for a few days after birth, and are then killed, and the parents necropsied and examined. Given the different lengths of time to become pregnant, this test can be as short as 5 weeks, or longer than 7. While the published database for this design is small, businesses and companies around the world are using this design currently. A large database should be available in a few years.

The NTP took an even more stringent approach, and reduced the time further, to 21 or 28 days, depending on which version is being discussed. At each dose level, this design uses one group of males, and two groups of females. One group provides information on developmental toxicity, and is dosed only during gestation; the pups are evaluated after birth for survival and growth to pnd 4, since most severe structural/functional problems become evident by that time. The second group of females is used to assess female reproductive function. They are dosed for the entire duration of the study, evaluating the ability of the females to ovulate, mate, and implant. After impregnating the first group of females, the males are dosed for the remainder of the study, and necropsied at the end. Because only a fraction of spermatogenesis is exposed to the toxicant, this design relies heavily on histopathology of the male reproductive system to correctly identify male reproductive toxicants. While this sounds straightforward, correct and informed histopathologic interpretation of the testis is still relatively uncommon. The shortcomings of this test are that it cannot detect occult genetic or functional damage in the cells that does not manifest as structural damage. In regards to developmental toxicity for this design, notice that exposure continues for organogenesis, but the pups are not evaluated for structural abnormalities per se. Instead, the emphasis is on alterations that threaten the animal's ability to grow.

Finally, the Chernoff-Kavlock test doses the pregnant female during gestation, and evaluates the weight and number of pups for the first 4 days post-partum. This test is short, and quickly identifies life-threatening malformations, or reductions in lactational ability.

Note that all designs are in active use: the OECD design is being used worldwide to generate some data on compounds that currently have no data available; the NTP design is being used on various projects (including one to prioritize drinking water disinfection byproducts for EPA), by the broad scientific community to help prioritize compounds for further evaluation.

## TRADE-OFFS IN USING SHORT TERM TESTS

1. The duration of such studies reduces from, say, 21 weeks (for the in vivo portion of the test) to 7 weeks, or even 4 weeks. This is a significant time savings.
2. What we gain in time, we lose in our confidence in a negative answer. That is, we are not sure that a compound that tests negative is really non-toxic. This was demonstrated with the NTP design. The authors of this design knew that it would be capable of identifying positives, but that the key question was: how sensitive would the test be in identifying a slightly toxic compound? They tested four chemicals of varying known reproductive toxicity (tested in the Continuous Breeding design), and found that, as expected, this short-term test missed one (the least toxic); there were no adverse effects seen for one of the chemicals. Additionally, by their nature, these short tests will not identify adverse functional effects on the second generation, or premature reproductive senescence.
3. When the "system is disassembled", each component needs to be evaluated separately. That is, short tests need to consider each component of female reproductive function individually (ovulation, fertilization, implantation, gestation, delivery, nursing). A corollary of this is that, for males, time limits on short term tests preclude the proper evaluation of germ cell mutagenesis, or spermatogonial renewal.

In essence, this strategy divides chemicals into two categories: known positives, and unknowns. Put another way: there will be compounds that have been shown to produce toxicity, and those that were not toxic in the short-term test, but that may produce toxicity when evaluated for longer durations in more thorough designs. This toxicity may be slight, but it may also work through a window of vulnerability that was not evaluated by the short design.

These tests can have other endpoints "piggybacked" onto them. The NTP uses the males from the 28 day study to provide hematology and clinical chemistry data, as well as histopathology on somatic organs of interest (liver, kidney, etc). Incorporating these designs into a short-term strategy that evaluates a wide variety of endpoints and systems should pose no problem.

## IN VITRO TESTS

In vitro tests are excellent for examining specific components of a process in isolation. For example, one can examine limb development in vitro and not worry about dispositional or detoxification processes interfering with the evaluation. They are also very appropriate for screening a group of compounds for a specific activity (for example, the ability of putative antibiotics to inhibit a bacterial cell wall synthetic enzyme). This use will be discussed further in the "New Strategies" section.

This very isolation is detrimental to a screening process. Good screens evaluate as much of a process at once as possible. Again, to cover in vitro what would be covered in vivo, multiple tests are needed.

Using male reproduction as an example, there are short (24-48 hr) in vitro methods for finding effects on spermatogenesis in vitro. However, to keep the cells alive, these methods are too short to correctly identify more than 20% of known testicular toxicants, they lack the testosterone-producing interstitial cells, and they lack the rest of the hormonal control systems (pituitary, hypothalamus). Thus, if there were going to be any confidence in the answer, this approach to male reproductive toxicity would require tests to evaluate those components of the system. The same is true for developmental toxicity and female reproduction: in vitro tests exist for some parts of each process, but not for all.

Since the overall process of reproduction and development is so complex, no "test-tube" assays have been evaluated as surrogates for in vivo testing. Receptor binding assays, second messenger tests, or other molecular endpoints miss so many of the potentially vulnerable processes that this attempt has not even been made, to our knowledge.

In short, this is a two-edged sword. Coupled with the lack of confidence that a negative answer in vitro truly means a lack of toxicity in vivo, we cannot recommend at this time the use of in vitro tests to correctly identify toxicants.

Structure-activity relationships would likely provide some clues, but work in this area in conjunction with developmental and reproductive endpoints is still nascent. Early indications suggest that it can find application to the broader areas of reproduction and development, but it is too soon to tell.

## REPRODUCTIVE TOXICITY AND GENOTOXICITY

It is theoretically possible that tests for genotoxicity would also identify reproductive and/or developmental toxicants. This hypothesis was evaluated using the NTP database, comparing the responses for a variety of genetox tests with outcome in the RACB test.

Overall, there are too few compounds tested in both systems to really evaluate the concordance, but generally, the results are not encouraging. Let us take the most promising relationship: if a compound was positive in the in vitro mouse lymphoma test, there was a 75% chance it would be positive in RACB. However, all compounds that were negative in lymphoma were also positive in RACB. So the sensitivity is reasonable, but the specificity is unacceptably low. Similarly, for in vitro cytogenetics, if a compound was toxic there, it stood approximately 70% chance of being toxic in RACB. However, if it was negative in cytogenetics, it still stood an 84°/0 chance of being toxic in RACB.

The preliminary indication is that we cannot hope that tests for genotoxicity will correctly identify reproductive toxicants. Based on more limited and personal evaluations over the years, our feeling is that the same is true for developmental toxicity.

## NEW STRATEGIES AND TECHNOLOGIES

To deal with the cascade of new chemical structures, new approaches are needed. Three can be recommended:

1. Use benchmark dose analysis. There is a single reported application to (male) reproduction, but several reports in the recent literature for application to developmental toxicity. The attraction of BMD is that one could use half the animals that are used in a definitive test, and the model would deal appropriately with the consequent (and slight) reduction in certainty, while yielding a approximately 40-50% cost reduction. Halving the animal numbers also halves the animal care time, the dosing time, necropsy time, tissue prep time, pathology time, etc., although it does not reduce overhead or various preparative costs associated with those activities or others. This also would not change the duration of the test. Although still relatively new, BMD holds such promise as to warrant it is being raised as the most likely solid improvement for this field.
2. Use a tiered approach to requiring information. This may involve some preliminary information triggering a request for further specific tests. This is the case with the new EPA Reproductive Toxicity Testing Guidelines: if changes in epididymal sperm count are found, a count of testicular spermatid nuclei is requested, both for confirmation and to identify site of effect.
3. Use SAR. If a previously-registered compound that reduced sperm motility is structurally related to a new candidate, requesting motility information on this candidate is a reasonable and targeted request. There are such huge benefits to be derived from computer-driven SAR methodologies that further work in this area is clearly warranted. The impact is biggest where the costs are greatest (which generally correlate with duration of exposure or numbers of manipulations of animals).

New technologies include the use of the computer for a variety of tasks: counting sperm and measuring sperm motion, counting and sizing ovarian follicles and stages of spermatogenesis (using image analysis techniques), and collating and producing an overall toxicologic profile. With the awareness that many transgenerational effects appear to result from the binding of xenobiotics to specific hormone receptors, one could imagine a screen of in vitro tests that assess the ability of a new compound to bind to a variety of hormone receptors and stimulate transcription. Such a vision has been proposed by others, with such receptor systems transected into cultured cells, so that the assays become in vitro cell culture systems. These are still in the planning stages, and it should be noted that, while simple in concept, they present significant technical challenges. Finally, transgenic animals are gaining acceptance as interesting model systems with some significant potential for application. While it is too soon to tell whether transgenics will be useful in identifying and ranking developmental/reproductive toxicants, we will note that many transgenics do have significantly reduced gametogenesis/fertility; whether this is a benefit or a drawback would depend on the question being asked, and the way in which it is being asked. This may be worth some additional consideration in the future.

## ▮ SUMMARY

Several strategies can be employed to significantly reduce the time and expense of preliminarily identifying reproductive and/or developmental toxicants. Each reduction in time and cost brings with it a concomitant reduction in certainty that a lack of toxicity over the short term also means a lack of toxicity over a longer exposure. Such tests are best used to prioritize compounds for further testing and evaluation. Benchmark dose and SAR strategies also can be viewed as valuable tools in the struggle to maintain public health at the least possible expense. If these reductionist strategies are not used to entirely replace longer, more definitive, tests, they can be used with confidence and success.