Interpretation of Low Incidence Findings in Reproductive and Developmental Toxicity Studies

Joseph F. Holson, WIL Research Labs, OH

DR. ELWELL: Our next speaker is Dr. Joseph Holson. He is the President and Director at WIL Research Laboratories in Ashland, Ohio. As a toxicologist he is well known in the field of reproductive and developmental toxicology. Today his presentation will be "Interpretation of Low Incidence Findings in Reproductive and Developmental Toxicity Studies."

DR. HOLSON: I don't want to waste time. So, I am going to begin. Pardon me, some of you, this maybe is excessively instructional, but as I look in the audience I see Dr. Christian is there, maybe four or five people with developmental backgrounds in the audience. So, I thought some comparison was in order to the oncogenicity area because the endpoints in these two disciplines of Toxicology are very important as irreversible efforts in Toxicology.

I know Don knows a lot about the repro area, but I think it would be good to step back a little with a couple of presaging comments here. I didn't get into power, although it is the crux of this entire issue, and the power of the study is dictated by the level of statistical confidence one wants, the magnitude of the difference between the treated and the control groups or the delta that you are trying to detest, the significance level in conjunction with the background variance of a particular measure or variable.

Now, the power of the study is inversely proportional to the variance. So, whether it be a single end point or a study done across laboratories these are hugely affected by power which dictates the necessary N before one is enabled to gauge the certainty of the final interpretation.

In the reproduction and teratology or developmental toxicology areas what is particularly important is the fact that because we know there is a high degree of concordance between human adverse effect scenarios on the fetus and what we can demonstrate in rodent models, including the rabbit a great deal of attention and significance is placed on low incidence findings of one, two or three.

Comparison of Overall Spontaneous Malformation Rates in Different Species

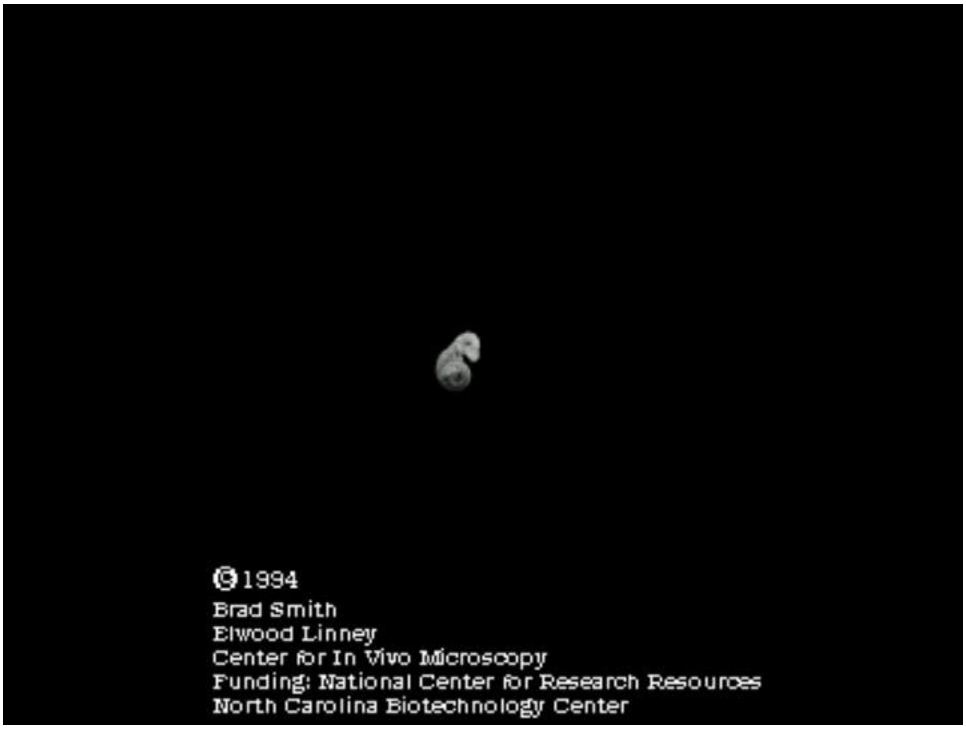
Species	Mean %	Range (%)	N
Rat*	0.33	0-1.6	9643
Mouse	1.2	0-3	5207
Rabbit	3.2	0-10	4708
Dog	5.5	5.3-5.7	167
Human	4.0	3-9	Multiple Surveys

* Actual number in laboratory population ~44,000

Additionally, concern for 2 or 3 malformations per group is appropriate because these will not be statistically significant from controls. This slide compares spontaneous incidences from multiple laboratory species and human beings. As you can see, malformations are far lower in laboratory animals than human beings. Why this is has not been thoroughly studied to my knowledge. But the shear infrequency of abnormalities in animal models as shown here demonstrates the basis of low-incidence interpretations problem at the regulatory review level.

Now, with that, I want to thank my colleagues, Ben Varsho who just changed the computer and Jeff Pitt, and Lewis Kaufman who did a lot to put our historical control data together for this presentation.

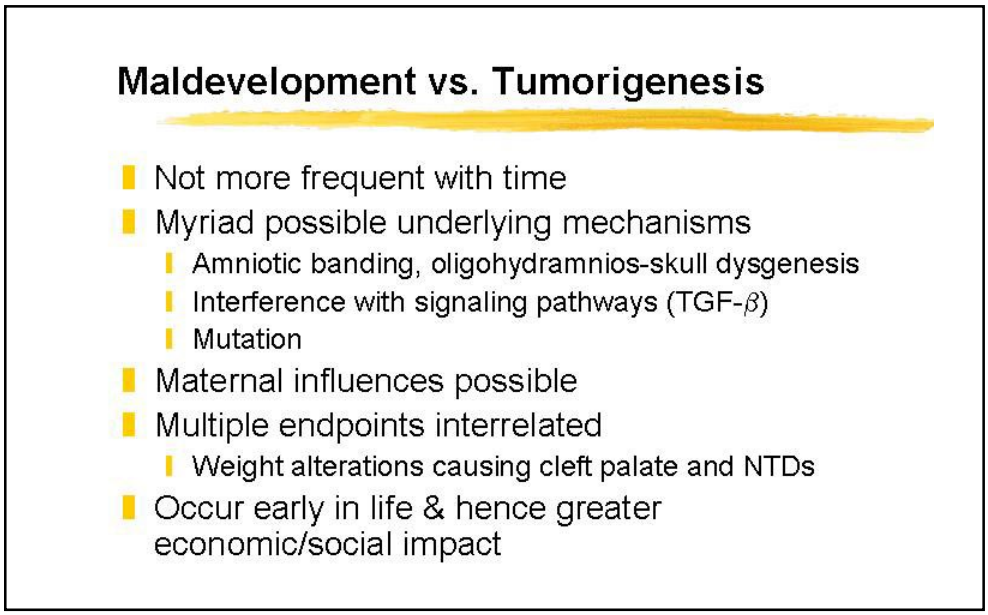
I am relying primarily on the data from the laboratory at which I work simply because it is done at one place. There are other databases. It is probably among the two largest in the world.



I am going to show you moving images from vital microscopy of an embryo that runs 5 seconds. It starts early in development, about 3 days after conception occurred. It is a mouse, and it is going to show you all that occurs in a 2-week period of development.

This is very important because we do in-utero exposure studies to look for developmental toxicity. It shows you how rapidly things are moving, and it points clearly to the fact that no hour, minute or single point in time is the same at any other point in development. Therefore any missed exposure, any hiccups in the course of that study and the conduct have great impact on its outcome.

This is done by Brad Smith, at the North Carolina Center for Biotechnology, and what you might notice as this scrolls through, you are starting from a very small embryo, as I said a day after implantation. You see the eyes, opening then closing, all of these structures forming in a few short 2 weeks.



Now, let me just drop back and look at maldevelopment which I am going to use as a general term during this presentation. One thing that is different from the cancer area is at birth we know an incidence of defects in humans.

By the time children are of school age we know that incidence raises to about threefold, but it stops then. You basically have ascertained the adverse total on development by that point in time, unlike in carcinogenesis where you can stage the long-term studies to look for tumor yield, and you know they are going to increase with age.

There is a myriad possible underlying mechanisms for maldevelopment, and it is amazing that the models work this well when you look at that, but everything from just plain physical phenomenon, a wrapping of an umbilical cord about a limb causing an in-utero amputation, too little water or urine being produced by the fetus leading to oligohydramnios and then skull dysgenesis is an example I am going to use with the ACE inhibitors, and interference with signaling pathways.

We now know there are 17 or 18 signaling pathways involving ligands which control a huge complex symphony of tissue-cell interactions, tissue migration, morphogenesis, a huge number of places that an exogenous agent may impinge on development, including mutations as in carcinogenesis.

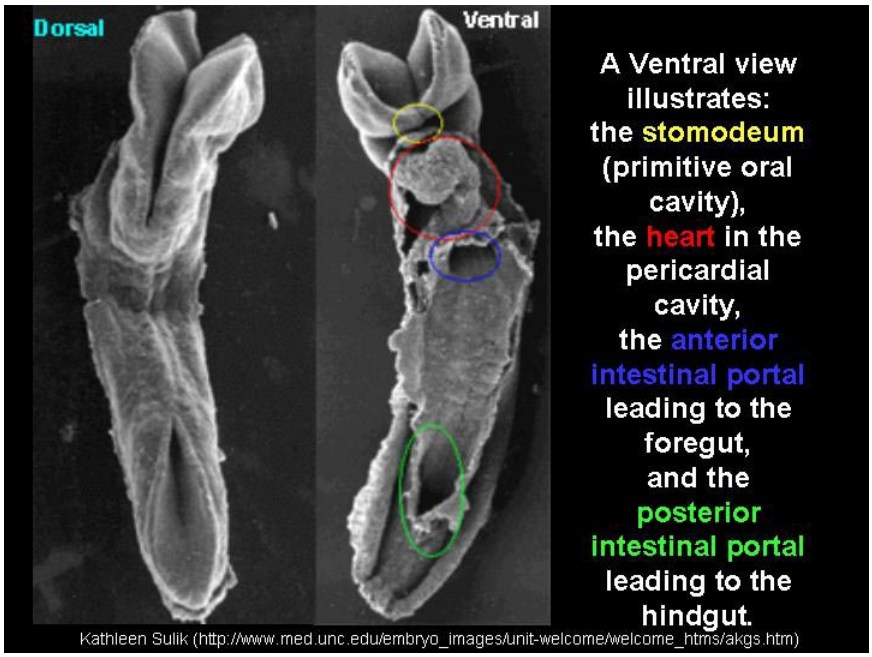
Maternal influences are possible. It is a too common misconception that maternal toxicity easily causes abnormalities, but certainly there are possible maternal changes and alternations that re-

sult in maldevelopment. In bioassay studies, maternal toxicity/potential weight losses in the 15-25% range frequently result in some fetal alterations as well as reproductive defects.

We are looking at multiple end points which are directly inter-related. Weight and body weight gain in carcinogenesis studies can have a certain context, but I want to show you how in this case weight or growth is extremely important to the embryo because any body cavity or structure beginning in an embryo is large relative to its surrounding tissue.

As development progresses embryos increasingly acquire the cortical masses, cavities diminish. You reduce the weight in a group of animals by just a small amount. You look at that same cavity subjectively, and it looks like a significant morphological change when really it was the result of growth retardation.

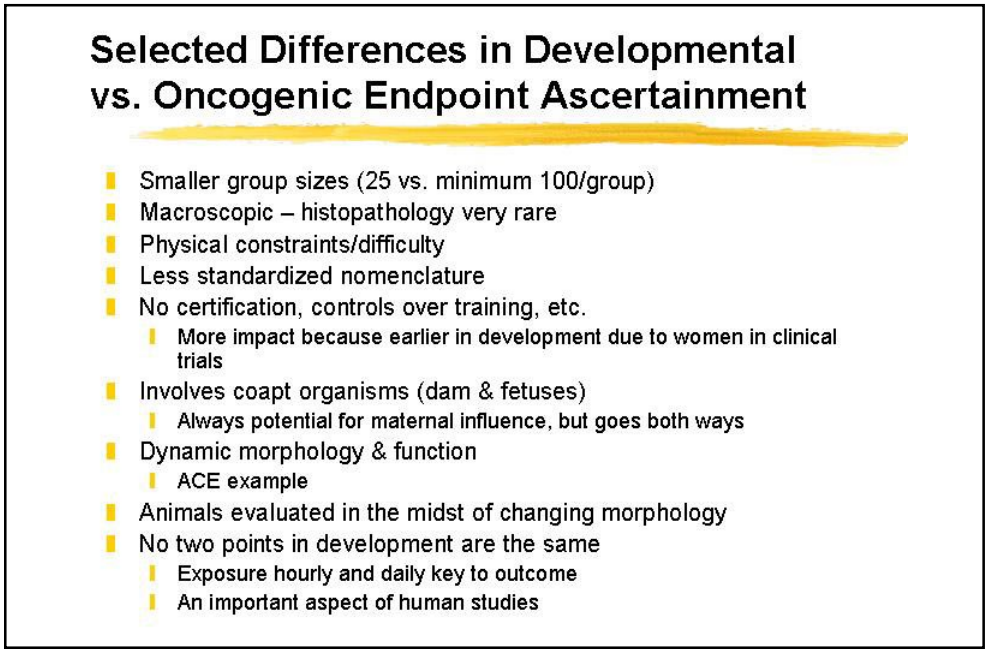
This may result in cleft palate and frequently in neural tube defects. These things occur early in life, that is birth defects from the human perspective very early in life and have great economic and social impact.



Now, to give you that example about exogenous influences here let us look at the dorsum of an embryo. This embryo is from Kathleen Sulik's lab in North Carolina.

Here you can see the spinal cord basically zippering up or fusing with the neural folds closing. This is anterior, and this posterior. When we turn it over and look you can see that this is the entire brain to be formed. Here is the stoma, the primitive oral cavity. Here is the heart appearing as bulbous tubular structures.

All of these are microscopic. At the time this picture is taken, the embryo is approximately 550 micrograms of wet weight, 98 percent water. You understand then that the small vesicles in which these embryos are growing and maturing could be very susceptible to any alterations in fluid movements or dynamics, electrolyte fluxes, etc., and here again showing you this fold here and the opening here of the anterior and posterior gut regions, but again to give you the feeling that at risk are influencing a lot of cellular interactions and tissue movements.



Because no prototype for this presentation could be found in the literature and the substantial interests in carcinogenesis of the present audience I thought it would be instructive to review similarities and differences between these areas. In smaller group sizes for one thing, 25 per group of animals, that being dams if it is a developmental toxicity study, sometimes in repro studies 30 or 35. There is generally a minimum of 100 per group in carcinogenesis bioassays.

The things we do to look at effects are largely, at least on fetuses all macroscopic done under dissecting scope, but still at the macroscopic level, not tissue subdivision histo-analysis at all, but there are lots of physical constraints and difficulties in judging normalcy in these specimens which I am going to show you.

There is a far less standardized nomenclature. There is an unofficial one published by Marta (MidAtlantic Reproductive Teratology Association) that has been published in the Teratology Journal, but not with the same rigor of standards or classification as in the carcinogenesis area.

There is really no certification or centralized training efforts or centers that typically imagine BS level individuals receive on-the-job training typically that are doing these critical fetal evaluations where one or two findings can either cause major labeling issues or kill a candidate product.

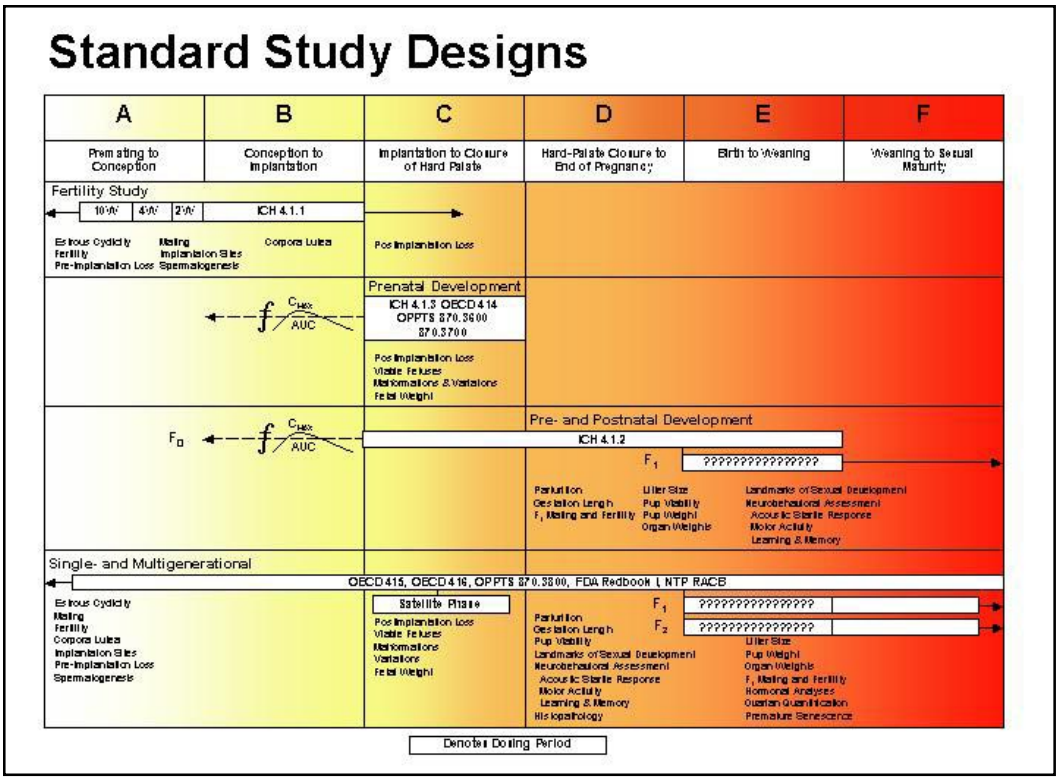
Now, that contrasts to people with veterinary degrees, extended education including board certification and having to learn a standardized nomenclature and process for classification. Further difficulties stem from the fact that in developmental toxicology coapt organisms, dams and fetuses are simultaneously exposed and evaluated.

Additionally, there is extensive interaction between the two. It is very difficult to separate the two. In fact, if you want to do a PK study, imagine trying to develop an analytical method that could measure the level of a chemical or drug in those 500 microgram embryos if you wanted to do a real PBPK model on a scenario of exposure in each pregnancy.

Evaluation of developmental well-being and outcome entails dynamic morphology and function. This is one of the most challenging aspects. I am going to give an example here. Morphology and function are progressing and changing. We are taking a snapshot look at them. We are not on the manifestation end looking at the resulting cancer, as a comparison.

No two points in development time are the same. Exposures hourly and daily can be key to outcomes. When we do metered dose inhalant studies for instance that is generally an hour or 2-hour exposure, and that is a small window in a twenty- four hour period of development.

What is important about that is when you try to do follow-up epidemiologic studies or birth defect registry studies, say, as a Phase IV or either under way to trying to take a product to market, those exposure differences can make all the difference between the outcome in animal studies and what one would find in the birth defects registry or epidemiologic study without ensured and similar exposure regimens.

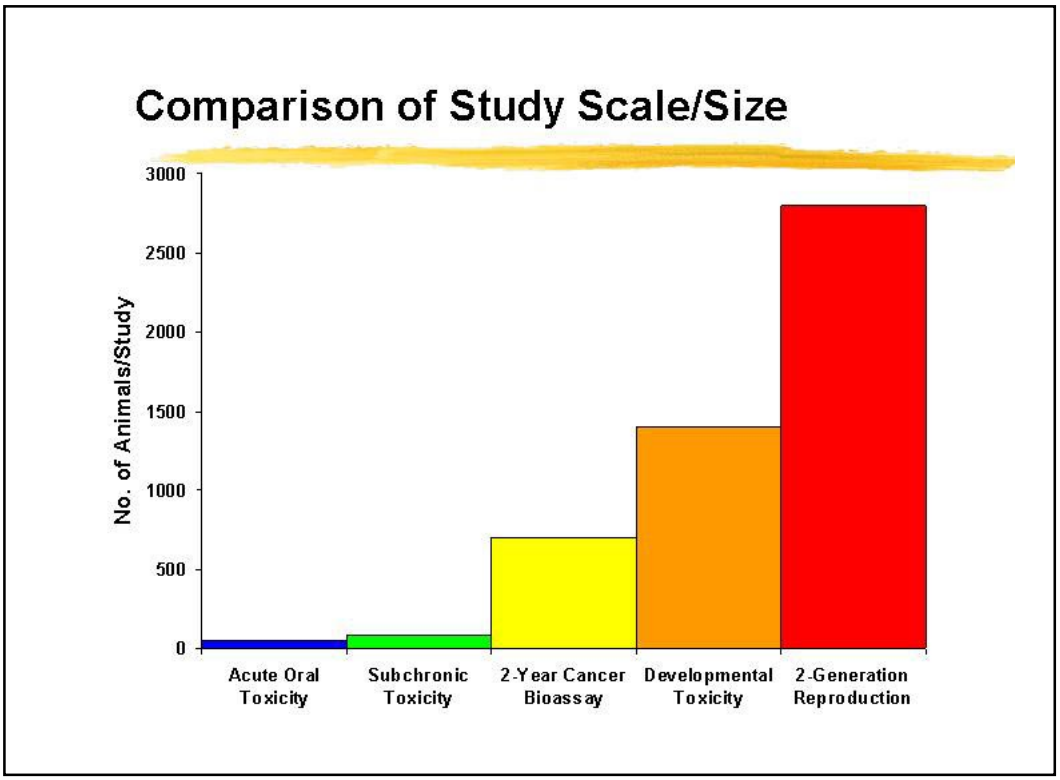


This slide is a little busy but let me just take your mind up here. I want to show you how we generate repro and developmental data. Here are major biologic events from conception to implantation.

This is the fertility arena. I am not going to address that today. These are the in utero figures, implantation to closure of the hard palate, just a convenient developmental anatomic event when the top of the mouth closes. It is discrete, but there is a lot of development differentiation occurring past that, the reproductive system, central nervous system, the lungs and the renal structures are maturing well beyond the time we typically expose, and into the postnatal period but these are the studies that involve examining the fetus primarily.

Then you go toward the latter part of pregnancy through postnatal life. This begins to involve these longer, larger reproduction studies. Down here the white line is demonstrating exposure.

This is a single or two-generation, multi-generation reproduction study. This will be the pre- and postnatal development study as provided for in the ICH guidelines supporting drug development which is less complete in terms of exposure compared to exposing through a second generation as in EPA required studies.



Another thing to take a look at is comparison of number of animals on the study. I have down here different types of studies, but take a look at a 2-year cancer bioassay study with somewhere up there around 650, 700 animals.

Look at a developmental toxicity study because you are taking each individual fetus from the litter, averaging 12 or 13, and you can see the huge number of animals here and then a reproduction study that is two generation going all the way up to around 2700 animals that are involved in the study must be evaluated or reported on.

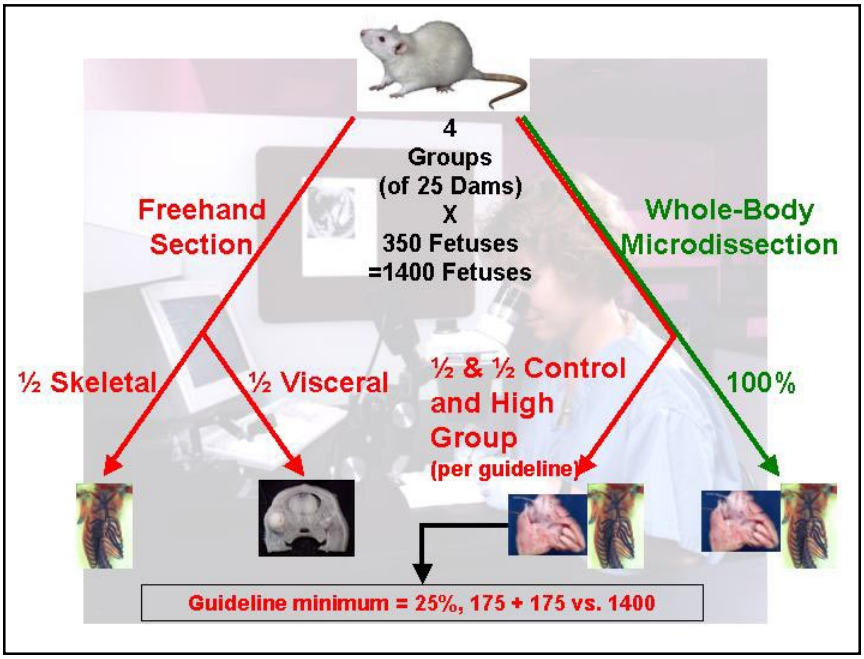
MARTA/MTA Database Fetal Abnormalities

Examination Type	Individual Descriptors
External	123
Visceral	277
Skeletal	467
Combined	**867**

www.hcd.org

Now, multiply that times the following. Here from the MARTA and MTA database of fetal abnormalities, just the in utero study this is the number of individual descriptors reported in those databases. Multiply that times the number of animals. Consider a providential level of 105? How many times should you see something positive, apparently positive? Quite a large number in these studies. This presents quite a challenge interpretively.

How do we obtain the data on fetal outcome? There are two basic procedures here and this has involved how you do the work, how you make the determination that drives a lot of the outcome as was just mentioned in the previous talk, but there is an old method that Dr. Wilson, my mentor had developed many years ago. It is called a freehand sec-



tioning method. You fix the fetuses. You slice them with a razor blade by hand every so many millimeters.

Guidelines allow you to do half the skeleton because in the Wilson method you can only do half by the skeletal examination, and you can only do one-half by visceral.

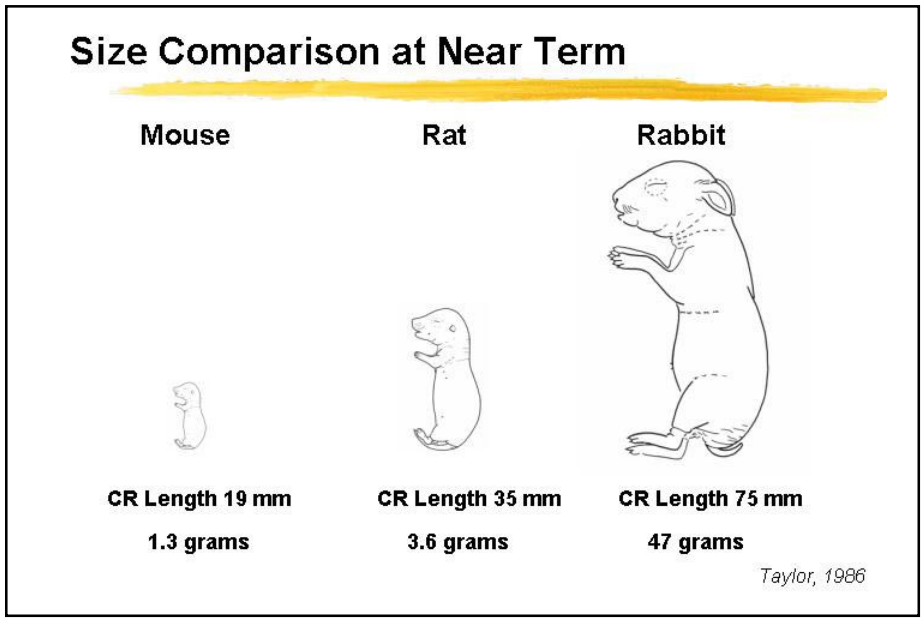
Now, the more preferred method these days is that of a whole body microdissection. Some refer to it as a fresh dissection. You basically remove the fetuses, they are not fixed before being euthanized and then an autopsy is done cervically to caudally. They are examined every structure for size, shape, color, etc.

However, ICH guidelines and the EPA guidelines allow you to do minimal here. They allow you to do one-half of the control and one-half of the high. If you do that you are really ending up examining 25 percent or 175 of the total of 1400 specimens.

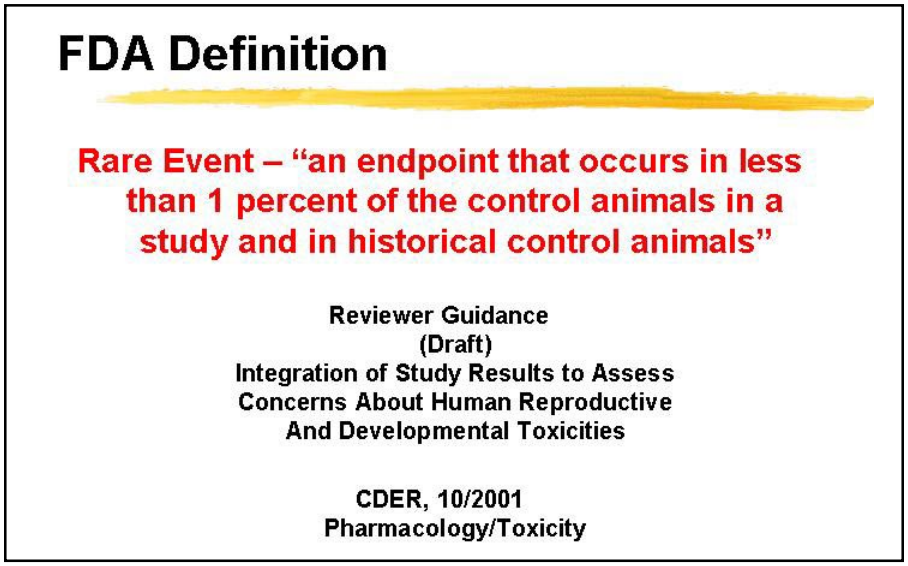
Now, I have never abided by this. I think it is one of the biggest mistakes that is done. I think it increases both the risk of false positives and false negatives, but certainly it increases the risk of dubious data sets because the power of the study is the crux of the issue in dealing with interpretation of these rare events.

What I am showing you here is just a fixed skeleton, a segment of a fetus here, and this is an open heart section, an open view of the heart looking for ventricular septal defects there.

So, you have this early decision of how you approach this aspect of study design. Most laboratories I know are currently doing the minimal here or something on the minimal side.

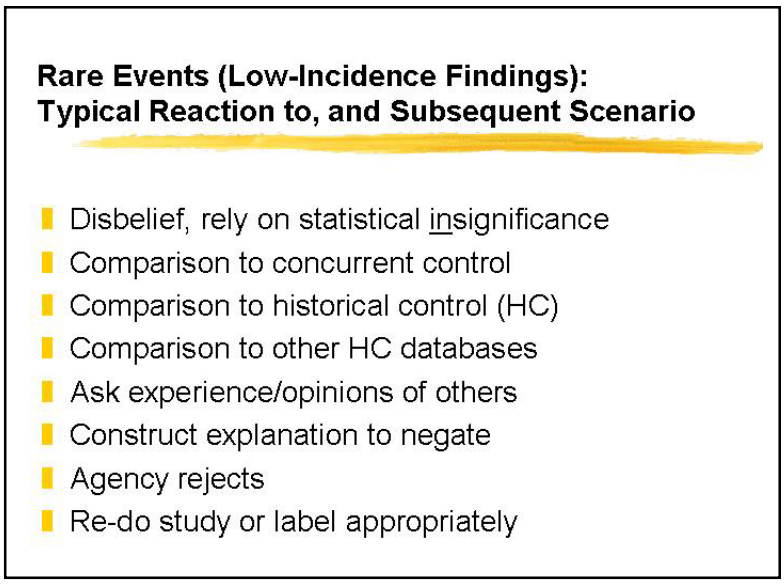


In addition to this issue consider the size of specimens examined, their small structures and take notice of the scale factor. Here is one species commonly used, the rabbit. This is the size. It is pretty easy to do this type of dissection procedure, but when you are doing a rat with a crown-rump length of about 35 millimeters, 3.6 grams. Then over here if you use a mouse it is even more challenging, 1.3 grams, 19 millimeters. When you look at the heart and have to make a cut through the heart valves to get into the heart there, you are dealing with a structure like a small pea.



One week or 2 weeks ago the FDA had an open meeting. It is the review of guidance relative to the new proposed integration of study results to assess concerns about human reproductive and developmental toxicities, given by CDER and the definition they quote in that document for a rare event is as shown here in red, an end point that occurs in less than 1 percent of the control animals in a study and historical controls.

I want to walk through this from some historical control data perspectives. My guess is what drove that, although I don't see Dr. DeGeorge or I don't know if anyone from FDA is here involved with that. My view is you will not pick 1 in 100 of anything up statistically in these studies, but you are going to see 1 in 100 far, far in excess of historical control backgrounds for almost all of the critical developmental defects.



This slide presents typical reactions to encountering rare effects.

I can say this because it represents my experience over 25 years of dealing with these issues. What happens when you run into one? Generally, it is just belief, intuition and you rely on statistical insignificance. That is it.

Compare it to the concurrent control. Compare it to the historical control, comparison of other HC databases and experience and opinions of others, construct some explanation to negate. Agency rejects. You redo the study or label appropriately.

This has been mentioned in particular in certain classes. NSAIDs if you look at the labels on that you will see it is in ACE inhibitors.

Classes of Reproductive and Developmental Toxicity

Reproductive	Developmental
Fertility	Mortality
Parturition	Dysmorphogenesis
Lactation	Alterations to Growth
	Functional Toxicities

In that same FDA guidance document there are categories of reproductive and developmental toxicities that are presented in the document.

What I am going to address today are those in the lighter green because I think they are the most perplexing, and I think these are the categories of effects that we most often see and have to contend with relative to rare events.

Selected Reproductive Endpoints Exhibiting Strong Signals from Rare Events/Low Incidence

Endpoint	Examples from WIL Research Historical Control in Crl:CD(SD)IGS BR	
Mean Viable Litter Size	13.9 ± 1.02	decrease of ≥ 1
Mortality ≤ PND 4	Mean = 96.2% Min/Max 91-95%	≤ 91%
Total Litter Loss	Mean = 0.94% (10/1061)	1 is equivocal 2 is more significant signal
Newborn Pup Weights	Mean = 7.0g ± 0.23 range 6.5-7.4g n = 1100 litters	≤ 6.5g strong signal

Next, I want to show you some selected reproductive end points that I have had experience with. This data again is from the lab at which I work, and I want to point some things out here that are very surprising.

I don't know how long ago the new IGS animal came out, about 4 years ago. At that time we were switching over from the old stock and they rederived the IGS stock. They were controlling the genetic pool by controlled random breeding in this new IGS animal.

So, we had done many studies, but since that time we collected somewhere around 9000 plus litter sets out of this, and this one mean viable litter size you see here, we have routinely either found in different compounds or drugs or have repeated multiple times studies where it was a questionable effect there.

Our experience in our lab with good husbandry, etc., and the data is consistent, a decrease of greater than one here or at one starts to present a real signal, and if crosses over about 1.25 it has in my experience been reproducible. These animals are very stable relative to that end point. It is one that wasn't taken as being that tight I think years ago.

The mortality here, a mean of 96.2 percent min/max 91 to 95 percent through the first 4 days of postnatal life. When you see anything at 91 percent or below it again represents a substantial signal of an effect. None of these show up statistically. Historically most labs would never make a call at this level. Some of these surprised me after I put the data together.

Total litter loss, this is a very interesting one, a mean of .94 percent. That means when a dam gives birth to progeny that she loses all of them. Now, in this case it is 10 out of 1061 litters. One is equivocal. Two or more is a substantial significant signal. It is unbelievable. You would never get any statistical test to demonstrate that. This has been reproduced time and time again.

Newborn pup weights, the mean 7 grams plus or minus 0.23, range 6.5 to 7.4, 1100 litters anything at or less than 6.5 begins to constitute a real signal.

Let me give you an example. This is an interesting one. This is the compound used widely over the counter pharmaceutically, OTC, required an inhalation study to get the appropriate blood level, area under the curve type profile in the study.

Case Study: *Dystocia, Extended Parturition and/or Pregnancy*

- 2-generation with second mating phase of F_1, vapor inhalation, used industrially, OTC pharmaceutically

PPM	0	70	300	500	700
F_0	0	0	0	2/24	3/26
F_1-1st	0	0	0	0	1/17
F_1-2nd	0	0	1/21	1/18	0/12

- HC then: 2/333 = 0.60%
- HC now: 4/1100 = 0.36%

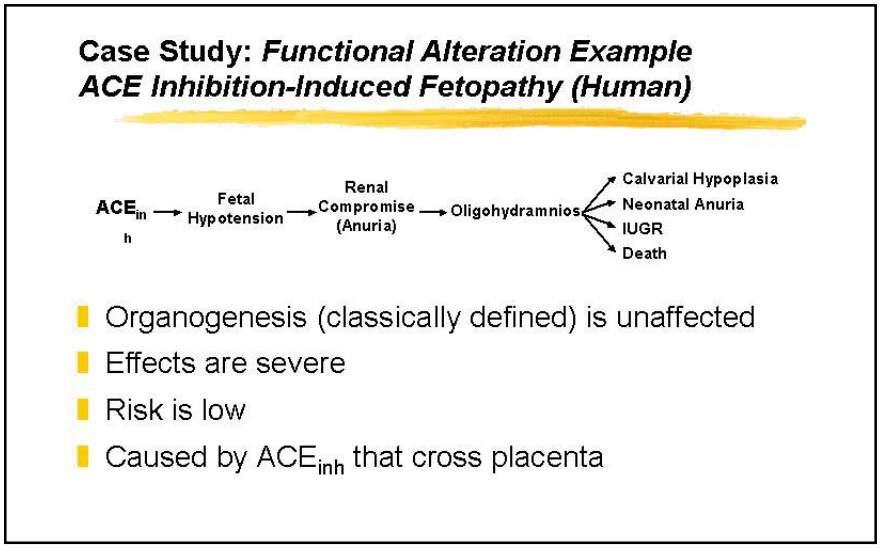
When we did this study we had a historical control in the IGS animal of about 333 animals. We did the study. You the F0, first filial or the parental generation. Out of 500 and 700 we had two and three dystocias.

Dystocia is a complicated or extended parturition where they show difficulties with the parturition process, and then we looked at the first generation. This fell off to 1 in 17.

Obviously what has happened here is as these didn't deliver, it changed the genetic composition of this group of animals. The first debate we had with the sponsor was, they said, "Well, when you do a two-generation study if you don't get the same effect in the second generation it must show that it doesn't count."

Not true. What this probably will often show us is an amplification of effect or imprinting because what is happening is you are removing those animals, maybe with the genotypic predisposition to the effect and then look what happened.

We came down to a second mating which we did to confirm this data. We came to a second mating, and it moved back down another exposure level, down to 300, again, the argument being this isn't consistent. It is not consistent because the denominator has changed and the composition of the gene pool in that sample of animals is altered.



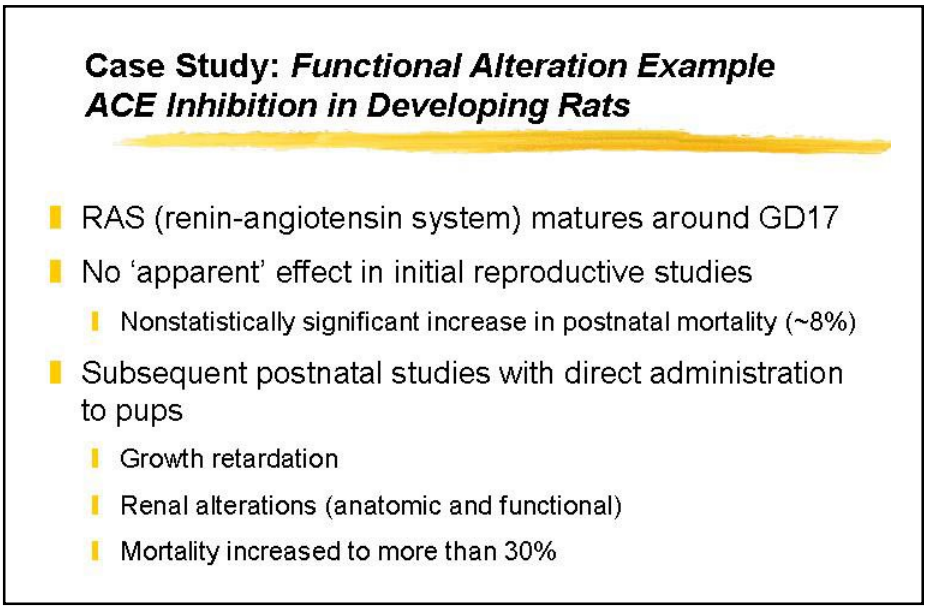
Another case study I want to show you is one of a functional alteration, a newer one. We have some other ones in the discipline, but this is the ACE inhibition induced fetopathy in humans first identified in humans although there was existing animal data, and it is an interesting story. When use ACE inhibitors, occurs in the third trimester, these antihypertensive agents cause fetal hypotension beginning at the end.

The effects consist of renal compromise in the fetus resulting in no anuria. That leads to a reduced volume of amniotic fluid, oligohydramnios. Because of this calvarial hypoplasia occurs.

The skull doesn't properly develop, because there is too much pressure without the buffering effect of the amniotic fluid accompanied by neonatal anuria, and intrauterine growth retardation. Obviously these babies are small and many die from this syndrome.

Organogenesis, that is that classic time from conception to closure of the hard palate is totally unaffected by these agents. These effects are severe. The risks are low, and it is being managed quite well through practitioners although there are still cases occurring from misuse, and it is caused only by the inhibitors that cross the placenta.

Why did this happen? When the original studies were done, and some people in this room may know about it, there was a lot of good work done in this area. There is a good review by Drs. Kimmel and Sonja Tobacova published in Reproductive Toxicology recently on this, but what it is is the renin angiotensin system matures in rat around day 17 or 18, and when they did the initial reproduction studies non-statistically significant increases in postnatal mortality occurred, but they weren't flagged, weren't heeded.



In subsequent studies the investigators reevaluate the incident well and determined that the absence of it involved direct administration in postnatal studies so that because the drug wasn't leaving the milk being bioavailable to the young rats while the renal matures postnatally because this is a developmental difference in timing between humans and rodents. In human beings the glomerular function in the kidney is maturing in utero not postnatally as in the rat. It was a case where the signal, the purpose of these hazard identification studies wasn't heeded because of the low incidence of postnatal death in the original studies.

Case Study: *Malformation Example*
Topical Antibiotic for Oral Mucosa

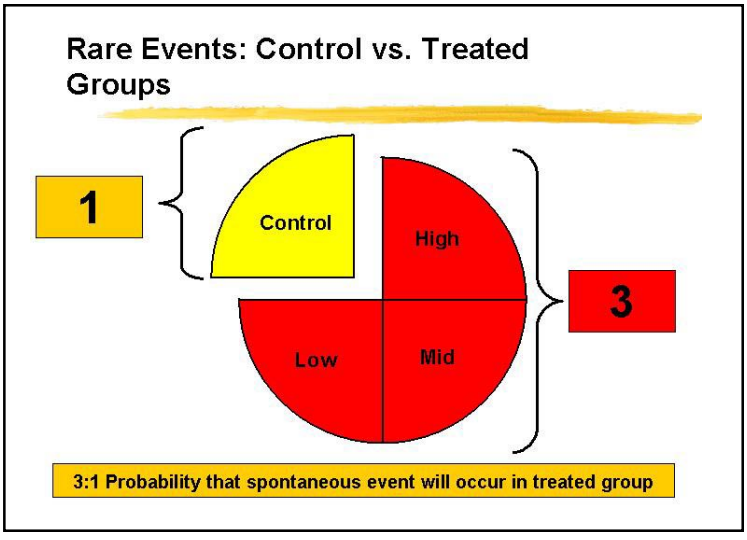
Rat Study Data				
Malformation	**1**	**2**	**3**	**4**
Retroesophageal Aortic Arch	0	0	1 (0.3%PL)	1 (0.3%PL)

Historical Control Data				
Malformation	**Total**	**Mean % PL**	**Min**	**Max**
Retroesophageal Aortic Arch	2/9643	0.02%	0.0% PL	0.3%PL

Another good example is retroesophageal aorta arch. This is a study that was done on an antibiotic to be used on the oral mucosa. This study, with absolutely no toxicity, no indication of any developmental effects, no perturbations in development in either of two species, and as you can see here we observed retro-esophageal aortic arch.

That is where the great arch of the aorta comes off the heart is behind the esophagus and it is very unusual. It requires a huge change in the patterning effects in early heart development, and you can see here observed one in each of one litter in both the mid and high dose groups. The historical control means you can see 2 in 9643. The minimum/maximum of 0 to .43 right on this now. With that rare event, 2 in 9643, how could you possibly have two in the treated groups, none in the control? Well, it is possible.

I can't prove it, but my view of this is because there is no other signal of developmental change and the nature of this drug in particular these findings were spurions.



I am not arguing that when we take historical controls that they don't represent a good sample of the population incidence in a laboratory, say, or in the whole stock of animals nationwide but what I will contend is that when spontaneous events occur in our standard designs you have got three times the chance they will

occur in the treated groups than in the control groups, and there is really in that study I just showed you, that is all they require is a 50/50 probability to have them fall like that. Because they are rare, however, it gives you great consternation in making those judgments and determinations.

Malformations which Have Occurred as Rare Events in Numerous Scenarios

Malformation	Incidence (%PL)	
	Rat	Rabbit
Ventricular Septal Defect	0	0.02
Cleft Lip/Palate	0.02	0.04
Abdominal Wall Defect Including Gastroschisis	0.04	0.06
Hydrocephaly	0.03	0.20
Spina Bifida	0	0.17
Renal Agenesis	0.01	0.02
Diaphragmatic Hernia	0 (2/39442)	0.04

Malformations can involve any tissue or structure and may constitute a rare event issue

I want to show you a list of malformations here which I think are fairly common in terms of appearing as rare events or low-incidence findings, and I want to give you the incidence here in our historical controls. So, you can see just how rare these are.

Here is ventricular septal defect in the rat. In our last nine thousand and six hundred fetuses and we have not had one. Now, I know of two companies whose laboratories will have 200 of these in a like database.

It is hard to explain, but remember the practical aspects of this which I described earlier, going in that heart, making the cut; it is very fragile. There are even studies by Dr. Solomon and others out of Glaxo-Smith-Kline showing that when they thought they had produced them in their lab and they allowed animals to live postnatally USD's were no longer there, either transient, a result of developmental variability maybe they were artifacts of dissection.

Cleft lip, cleft palate .02 incidence percent per litter, .02, .04 in the rabbit, abdominal wall defects. We had one chemical that we worked with and reproduced five studies working in this range spread across even would not increase with dose but every time we redid the study we kept getting them in the treated groups, in all groups.

This is what even in humans occurs in clusters quite often. Hydrocephaly 0.3, .20 in the rabbit. Spina bifida 0, 0.7. Renal agenesis .01, .02. Diaphragmatic hernia 0 in our population of nine thousand and six hundred plus.

When I went back and looked at the previous 8 years or something we had 2 out of 39,442. So, what does that tell you? The ascertainment of it, these are indeed very low incidence findings and the size of your historical control database is going to be a big determinant of how effectively the problem may be dealt with.

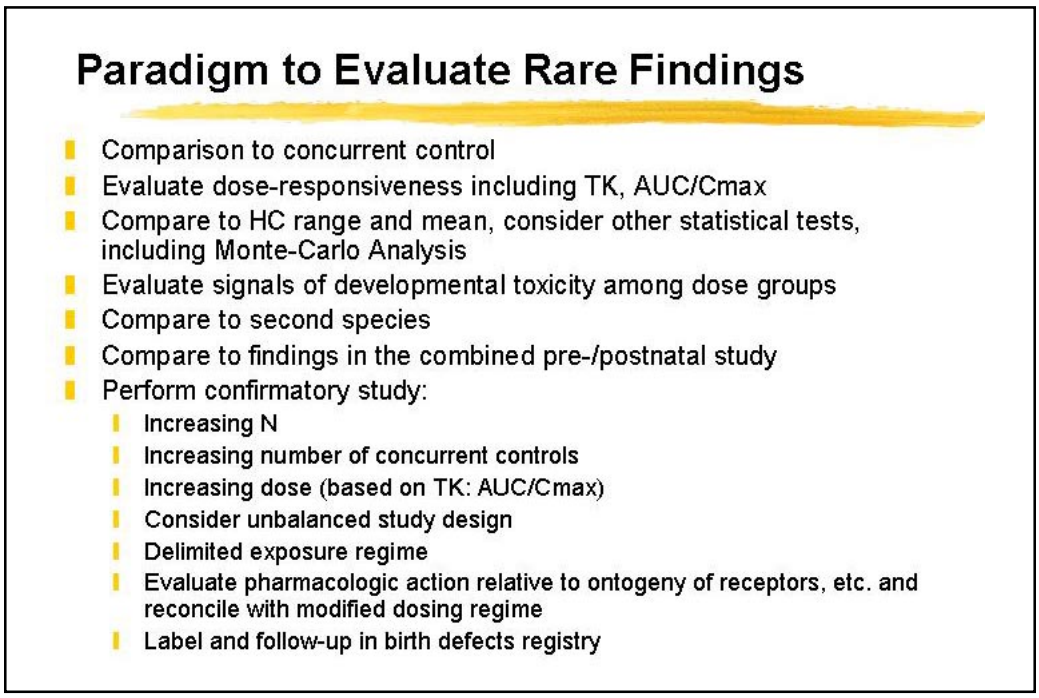
Rare Event Manifestation Matrix

Control	Low	Mid	High
0	0	0	1
0	1	0	0
0	0	1	0
0	0	1	1
1	0	0	1

Now, just take a look at this slide for a moment, I have seen these combinations. There is a big difference in how they do occur, too. Most of the time in the non-clinical area we can tolerate this although it could still lead to labeling consequences, but you use such a high dose to give you some kind of a margin through your therapeutic index or AUC or C max multiples in the development program.

When it occurs here it is a real problem if it is rare enough, but again it is really hard for me to be convinced with an incidence of one, whether or not it is in a human population, unless you know of a class effect or some other weight-of-the-evidence data. One observation associated with a new chemical entity is very tough to make a call on. I typically wouldn't.

Here is another one, in mid-dose group, another set of problems, but this case is in the retro-esophageal aortic arch. It adds a lot more weight because you are getting the two but still on a probabilistic basis that can occur at a 50/50 chance and then this one is interesting to me how often people ignore this. If that is a signal, this could be a signal equally as well when you are dealing with something at .02 percent incidence.



So, to attempt to be constructive I put together this paradigm, and I thought this is the way that I and other's look at evaluating rare effects or low incidence findings. Comparison to current control is always the first effort and the agencies tend to like to heavily rely on that comparison, but with the limited power in the studies we do it is a tenuous process.

Evaluation of dose responsiveness relative to internal dose a relative to area under the curve with the C max levels, compare to the historical control. Often, the FDA argues the comparison must be made to the mean not the range of HC values. I think that is too austere because the HC database is more of an estimate of the relative mean as compared to the concurrent control especially a sample estimate.

You might consider other statistical tests, including the newer versions of the Monte Carlo analysis, because it is really the only thing we currently have that is amenable to looking at sampling error and predicting population estimates relative to sampling values.

Evaluate other signals of developmental toxicity. Compare to second species, also, with the TK data, compare the findings in the combined pre- and postnatal study. Since we do have three segments you can also try to match if your doses or you AUC data is available, you can look at what happens in the postnatal animals.

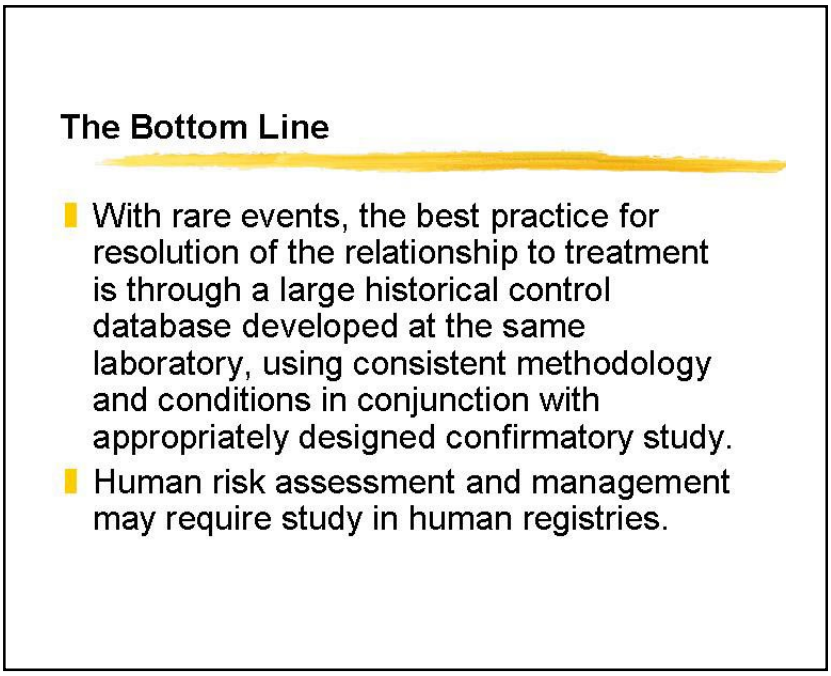
Do you think you see VSDs? Were they found at all? Do you have mortality that could be a result of VSDs, other effects in the postnatal period? So, make use of your total study set and perform a confirmatory study.

In my experience it is almost impossible to avoid a confirmatory study when you have exhausted the other things that I just mentioned, but remember this. Consider increasing the N. Increase the number of concurrent control groups. That hedges against the probability of three or four to one experimental to control.

Increase the dose based on the TK, AUC or C max estimates. In the confirmatory studies unbalanced study design is one of the biggest possible attributes. We automatically work against ourselves with standard balanced designs because as dose goes up power goes up. As response goes up, if you have an effect, power is less diminished. So, loading more animals on the bottom side of that curve is more effective in providing study data that mitigates against spurions low incidence findings.

Consider a delimited exposure regime because no point in development is the same. If you know the developmental timing of the organ or system effected, you can go back and just administer doses giving you the same AUC on the limited regime which could illuminate whether or not the underlying embryology and the outcome make sense.

Evaluate the pharmacologic action relative to ontogeny of receptors, etc., and reconcile with a modified dosing regime. Label and follow up possibly in a birth defects registry.



The bottom line on evaluating and reading rare events is that the best practice for resolution of the relationship to treatment is to move a large historical control database developed at the same laboratory that conducts the study, but you have got to have that database internally for the end points you are judging using consistent methodology, conditions in conjunction with an appropriately designed confirmatory study.

You are not going to do it mathematically alone. No one has moved the odds advantagally with the Monte Carlo analysis that I know of but I think it could be used more frequently in the future, and then human risk assessment and risk management may require further a study in human registries. Likewise, these are very limited in their resolving power.

Thank you.

(Applause.)

DR. ELWELL: I think we have time for one or two questions.

Dr. Kluwe?

DR. KLUWE: Thank you, Joe, a very nice talk.

Bill Kluwe, from Pfizer. I think those suggestions you made on following up some of the low-incidence effects are extraordinarily insightful. In reality though they work fantastically if those follow-up studies then confirm that that event actually did occur, if you can repeat it.

In your experience are they nearly as useful when in fact you do those studies and you can't confirm that? Does it get rid of the previous event or is it still there as sort of a black mark?

DR. HOLSON: That is a good question. It depends and varies with your relationship with the regulatory officials perhaps.

(Laughter.)

A negative confirmatory study must always, be definition, mitigate against a previous positive finding to some extent. The extent depending on the quality of the confirmatory study. This is based on weight of the evidence considerations.

DR. HOLSON: The major problems is that not enough of these scenarios are made public and hence, to heuristic value is lost.

You know, you have got to take the view that there is more sensitivity for these issues amongst regulatory people than there was in the past.

In general, if you do a robust confirmatory study, under these conditions, I think it should remove doubt, maybe make you still go with a registry study and do some post-marketed follow-up.

Tom, how are you?

DR. MARKS: Tom Marks, AstraZeneca. Joe, there is no question we can get positives. We have got over 1000 positive teratogens.

DR. HOLSON: Tom, that is a major mistake and erroneous statement.

DR. MARKS: Okay.

DR. HOLSON: There have been reported 1000 such "teratogens" at MTDs. In the concordance studies conducted, there aren't but approximately 40. There may be 1000 if you blindly ignore the fact with that of more than 600 of them humans were never exposed. It makes only non-sence to purport such interpretation of our total effort in this discipline.

DR. MARKS: Okay, conceding we have no problem getting positives, I would say 10 percent of the compounds I have worked on have come up with some kind of positive, now if we start taking rare events and we look at this new guideline that the FDA or guidance that the FDA is putting out we are going to see more and more positives, but we can go back and say, "Well, we have got positives for aspirin.

We have got positives for vitamin D. We have got lots of positives." How do we, and we can go back and look at ACE inhibitors, for example, and say, "Well, yes, we should have picked that up, and we can show retrospectively we should have," but how do we determine when we get these rare events which ones are actually going to be hazards to the human or not?

We have got lots of them. I can give you a study that we are working on right now and it has got these rare events in one animal. So what do we make of this?

DR. HOLSON: One animal, rare events.

DR. MARKS: The rabbit and the monkey.

DR. HOLSON: Yes, well, if you do monkey studies like most people do, you are lucky not to have a rare event because your N is so out of sync with anything that is statistically effective. That is a real big problem in that particular species. Also, many people overuse this species because they are blinded by plylogeny.

Now, I, personally, think and maybe Dr. Christian and others here I recognize that one of the things that has happened here is we all have drawn too sharp a line in the interpretation of the observations.

I hear what you are saying, but recognize that most of the companies are doing pretty well in developing new products. This is not stopping new products. The labeling, the flexibility of the agency relative to labeling I think has been quite good.

I think it is important to label specifically, and I think it is important to improve our moxy relative to the preclinical conduct of studies along with postmarketing follow-up. It is really the only solution, but these studies are good for hazard ID screening out the bad ones. At least that is my experience with it anyway.

MR. VARSHO: Varsho. Can you say something about the overall spontaneous malformation rates comparing say some of the animal species to the human?

DR. HOLSON: Yes what is interesting is if you compare them the different ranges up to 100-fold. In humans, and we don't know why, but if we consider that every specimen from a non-clinical study is evaluated in minute detail and babies are born with minimal examination the difference becomes even more dramatic.

We don't x-ray all newborn babies. Even with that human incidence is about 4 percent. Compare that to the rat and the mouse. It is somewhere between 1/4 and 1/10 the incidence.

Now, what I am telling you is if it is because experimental animals in laboratories are not exposed to so many agents it is a bad sign because it means the human is exposed to agents that are doing it. I don't believe that.

Now, the other thing to remember is human loses about 30 percent of all conceptions, very early and that is to weed out defects but yet the rates are still high.

In animal studies you have a very, very low background incidence, and Tom, that is one of the key points. If power through study design is not maintained incidences of 2 or 3 observations become apparently treatment related and are difficult to explain. Rodents are quite consistent across the board relative to spontaneous defect rates, but in humans rates are much higher.

Bill? (Dr. Alaben, FDA)

PARTICIPANT: Yes, Joe, very nice presentation. In the one slide you mentioned that when you looked at the historical data, control data at WIL on septal defects you said that you had zero incidence of that but when you went outside the laboratory to other laboratories; I use the plural there presumably, you said that you had an incidence of I think you may have exaggerated but 200.

I mean can you offer any explanation about the differences? Are there strain variations in this? Why such a variation?

DR. HOLSON: It is a good question. I think it is probably a place I wouldn't want to go, but I think most everyone around the world is using the new IGS rat, in Japan; I have looked at data from various places. I feel a lot of it is the criteria being used by the people making the calls and in conjunction with dissecting difficulty. The dissection is not easy. Also remember VSD's likely represent incomplete development and may dose postnatally. We do not know enough to answer this yet.

DR. HARRIS: Jane Harris, BASF, I recently had a spate of, well, it was one or two studies with an increase in agenesis of the kidney, you know with one kidney and it was very interesting, you know. It was actually a low-dose response that decrease, and my company ended up doing a lot of rabbit. This is rabbit and ended up doing a lot of rabbit repeats in answer to that question.

You know, the issue is this. How do you negate something that you have seen, and short of vitamins and hormones you very rarely see an inverse dose response. So, I was suspect to begin with. So, now

we have to go to the agency and present this data, and I am really concerned about it because I don't know how they are going to respond to something like this.

DR. HOLSON: I don't blame you. I would be concerned, too.

(Laughter.)

DR. HOLSON: If you have done good confirmatory work with so many parameters it certainly is going to help relieve their concern, and it is a difficulty, and it is not different than in the carcinogenicity area but –

DR. HARRIS: But what is so interesting is since I have seen that effect I have seen more and more agenesis of the kidney popping up in almost every study, at least one if not two in every study I have done to date since that period which was going back a couple of years ago. So, I mean things change.

DR. HOLSON: They do change. Our experience thus far with the numbers we have in this IGS stock is that they are fairly consistent. We need to be watching that. We need to be communicating and reporting our databases to pick these up, but I will tell you this other thing that sometimes creeps into the picture. Somebody sees something. They start looking at it close and its incidence increases.

I recently had a data set sent to me for review, it had 30-odd club feet in the rats at the high dose, 15 club feet at the mid-dose, 5 in the low dose and 6 in the controls. You know that out of 50,000 fetuses how many club feet we have had? Zero. I have never seen one in my life. That is a miscall and misnomer, and some of these are subjective and sometimes you have just got to lay it on the table and say, "I don't believe the call," but they will creep in number as people get familiar. Subjectivity and technical training, experience and oversight are key contributors here.

DR. ELWELL: I think we need to continue our questions.

DR. HOLSON: Milly had her hand up. Let her speak.

DR. ELWELL: Did she really have it up?

DR. HOLSON: We don't have many developmental people in here.

DR. ELWELL: How many people saw it up? Okay.

(Laughter.)

DR. CHRISTIAN: This is just very short and you are most fortunate I don't have a voice today, too, Joe, but I thought you did a really good job, and I just wanted to emphasize one thing that you said which often is missed, and that is in looking at these rare events how important it is to put the whole picture

together and look for related effects and dose-dependent effects in the other end points that go together, and Joe said that, but it is exceptionally important.

These things don't occur all by themselves when they are usually associated, and there is a real causal relationship, and Joe said it. I will repeat it. Remember that.

DR. ELWELL: Thank you, Joe.