

---

# New Perspectives in Modeling of IR Induced Carcinogenesis

Igor Akushevich<sup>1</sup>, Galina Veremeeva<sup>2</sup>, Anatoly Yashin<sup>1</sup>

*in collaboration with KG Manton<sup>1</sup>, A Kulminski<sup>1</sup>, M Kovtun<sup>1</sup>, J Kravchenko<sup>1</sup>,  
AV Akleev<sup>2</sup>, S Ukraintseva<sup>1</sup>, L Akushevich<sup>1</sup>, K Arbeev<sup>1</sup>*

<sup>1</sup> *Duke University, Durham, North Carolina, USA*

<sup>2</sup> *Ural Research Center for Radiation Medicine, Chelyabinsk, Russia*

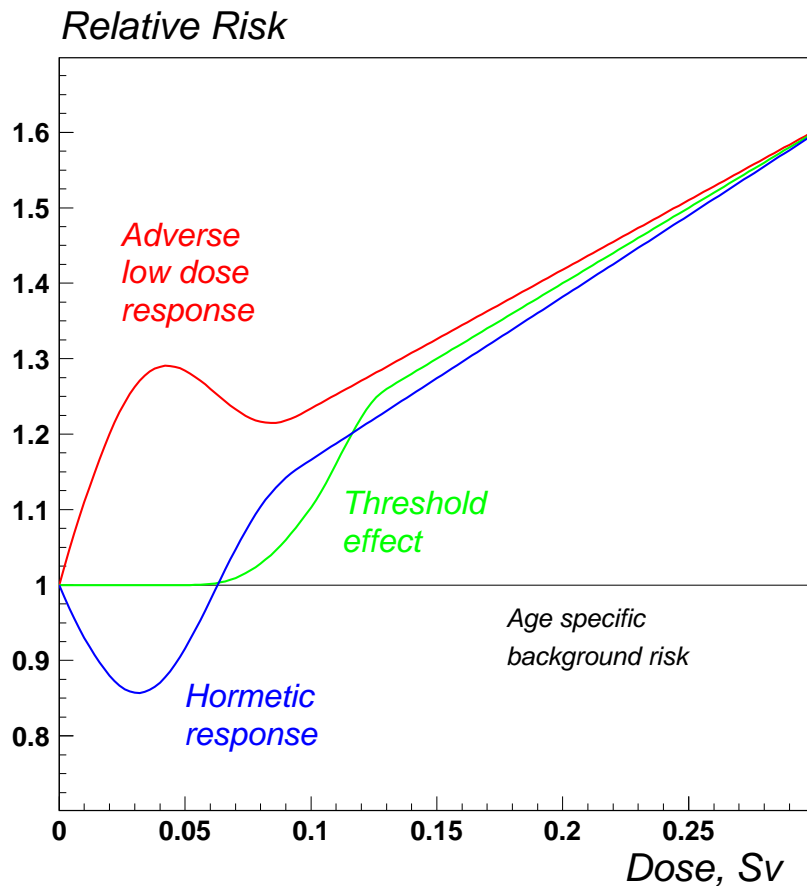
# Why and when we need population or bio-molecular models

---

- Analysis of crude data is not enough to make conclusions about the phenomena of interest. A common situation is when general dose trend can be estimated for the total, but not for stratified population (or cohort).

# Why and when we need models? (cont.)

- Investigation of specific biological mechanisms responsible for health effects of ionizing radiation. Assumptions of standard analyses (e.g., linear dose-response function) are often not valid



Nonlinear models of relative risk

## Why and when we need models (cont.)

---

- Analysis of crude data is not enough to make conclusions about the phenomena of interest. A common situation is when general dose trend can be estimated for the total, but not for stratified population (or cohort).
- Investigation of specific biological mechanisms responsible for health effects of ionizing radiation.
- Forecasting, i.e., projection of health changes and associated medical expenditures. Playing “what-if” scenarios. Investigation and development of different preventive and interventional strategies.
- Data analyses when data have non-standard structure, e.g., categorical data measured longitudinally; tracking event data; gene expression data; complete information from death certificates; datasets combining several sources; missing data.

# Types of datasets in radiation epidemiology

---

- *Data on case-control studies*
- *Grouped data, i.e., PYR, number of cases*
- *Epidemiologic Register*
- *Follow-up with covariate measurement at regular time intervals*
- *Follow-up with covariate measurement at irregular time intervals*
- *Tracking of individual medical histories*
- *Surveys, i.e., categorical measurements*
- *Follow-up of categorical measurements*

In many cases data are collected in an irregular study design. Data often include a combination of several sources and in the majority cases they have missing data.

In such situations, standard methods often fail and a comprehensive analysis becomes possible only by using mathematically appropriate models.

# Modeling vs. Empirical Analysis

---

Advantages of using biologically motivated models vs. empirical analysis

- Allow for estimation of biologically motivated parameters;
- Attract information on mechanisms of carcinogenesis;
- Do not contain implicit assumptions of empirical analysis (e.g., on proportional hazards);
- Allow for estimation of baseline carcinogenesis models;
- Allow for elaboration of clear strategies for sensitivity analysis;
- Have the potential to be combined with models of survival of cancer patients;

Two types of models

1. **Population Models**, where statistical unit is an individual
2. **Bio-Molecular Models**, where statistical unit is a cell

---

*Part I:*

***Population Models***

# Brief Review of Population Models

---

- Earlier population models: *Gompertz (1825), Strehler and Mildvan model (1960), Sacher and Trucco model, (1962);*
- Frailty models; correlated frailty models; debilitation models; repair capacity models (*reviewed by Yashin et al, 2000*);
- Quadratic Hazard Models (*Woodbury, Manton 1977; Manton et al., 1992*)
  - *Health state is described by a set of risk factors or covariates;*
  - *Hazard (e.g., mortality or cancer incidence) is a quadratic function of covariates;*
  - *Allows to model jointly dynamics of risk factors and mortality;*
  - *Structure of data: follow-up with measurements of risk factors at regular time intervals;*
  - *Mathematical formalism: stochastic differential equations or Kolmogorov-Fokker-Planck equation;*
  - *Recent generalizations (Yashin et al., Math Biosci. 2006) allow to include the effects of “optimal” (normal) physiological states and allostatic load.*

# Brief Review of Population Models *(cont.)*

---

## ➤ Stochastic Process Model (*Yashin, Manton (1997)*)

- *Generalize quadratic hazard model;*
- *Uses joint likelihood for estimation of mortality and risk factor parameters;*
- *Automatically generates values of risk factors to fill missing data;*
- *Time between surveys may be not fixed.*
- *Appropriate model for analysis of longitudinal data with irregular measurements.*

## ➤ Microsimulation models

- *Simulate individual age trajectories; population characteristics are calculated by averaging of the individual trajectories;*
- *Work with complicated dynamic model (e.g., nonlinear, history);*
- *Allow to construct projections for various “what-if” scenarios (e.g., Akushevich et.al., Risk Analys. 2007);*
- *Have the potential to investigate medical interventions;*
- *Can be joined with quadratic hazard models (Akushevich et al. 2005).*

# Brief Review of Population Models *(cont.)*

---

## ➤ Latent Structure Models

- *Examples: Latent Class Models, Grade of Membership, and the recently developed Linear Latent Structure (LLS) Analysis (Kovtun et al., 2007);*
- *LLS allows to work with high dimensional categorical data;*
- *Simultaneously investigate properties of population and individuals;*
- *Mathematical formalism: theory of mixtures and mixing distributions.*
- *Possible applications: analysis of gene expression data and biodosimetry.*

All of the above mentioned, and several other, models were recently reviewed by *Akushevich et al., (2006) in Radiat. Biol. Radioecol. 46, 663-674.*

# Perspectives in Applications: ETRC and TROC data

---

- Extended Techa River Cohort (ETRC) includes about 30,000 individuals exposed to protracted IR as a consequence of released radioactive waste into the Techa River during the initial years (1949-1956) of operation of the Mayak nuclear facility (weapon grade plutonium production) in the Southern Urals region of Russia near Chelyabinsk city (*Kossenko et al., 2005, Radiat. Res., 164(5): 591-601.*).
- The information on individuals in the ETRC is linked to the cohort of first and second generation offspring (Techa River Offspring Cohort, TROC).
- The maximum cumulative doses to red bone marrow (RBM) reached 2 Gy and were mostly accumulated during the early years of exposure (i.e., 1949-1956).
- The exposure was of a combined nature, consisting of external gamma IR and internal IR mainly due to  $^{90}\text{Sr}$ . Early exposures continue to have effects because  $^{90}\text{Sr}$  is a long-lived radionuclide (the half life is about 30 years) and because  $^{90}\text{Sr}$  is incorporated in bones since its chemical reactivity is similar to calcium.
- ETRC/TROC represent “natural” unselected population.
- There exists longitudinally measured health state characteristics, internationally verified reconstructed whole-body and RBM doses, and an ongoing collecting of blood samples from about 2,000 individuals of the ETRC.

# Perspectives in Applications: SPM model and ETRC

---

The ETRC data are longitudinal, i.e., there are repeated measurements of health status and other indices resulting from medical examination. Time intervals between measurements in ETRC are not fixed, and there are missing data.

Therefore the appropriate model is the Stochastic Process Model (SPM), which is capable of simultaneously describing dynamics of health indices and hazard rate as a function of the indices.

How can IR doze (or dose power) be incorporated into the model?

**Way 1** *IR dose can be considered an independent covariate.* Specific assumptions of coefficients in dynamic equations have to be made in order to reflect the averaged dynamics of IR dose accumulation in the human body conditional on the health state, reduction of incorporated radionuclides during later life, and the decline of the external exposure.

**Way 2** *Model parameters can be assumed to be dose or dose rate dependent, e.g.,* hazard function can be assumed to be linear in respect to dose which follows from the currently used hypotheses of standard empirical analyses.

# Family Analysis and Correlated Frailty

---

**Hypothesis:** There exists a correlation between the excess of relative or absolute risks of exposed populations (ETRC) and their offspring (TROC).

**Modeling strategy:** Model marginal bivariate survival function  $S(x_1, x_2)$  by averaging conditional one  $S(x_1, x_2 | Z_1, Z_2)$  over the individual frailties  $Z_1$  and  $Z_2$  assuming that these frailties are two gamma-distributed random quantities with the two mean values equal to 1, variances  $\sigma_1^2$  and  $\sigma_2^2$ , and correlation coefficient  $\rho_z$

$$S(x_1, x_2) = S_1(x_1)^{1 - \frac{\sigma_1}{\sigma_2} \rho_z} S_2(x_2)^{1 - \frac{\sigma_2}{\sigma_1} \rho_z} \left( S_1(x_1)^{-\sigma_1^2} + S_2(x_2)^{-\sigma_2^2} - 1 \right)^{-\frac{\rho_z}{\sigma_1 \sigma_2}}$$

**Advantages:** i) no assumptions about parametric form of underlying hazard required; ii) informative consistency between univariate and bivariate survival models; iii) can be used for evaluating mortality and longevity limits; iv) can be generalized to include observable covariates extensively measured in the ETRC and the TROC and/or to include dose dependence of parameters into the model

**Refs:** *Yashin, Iachine 1997, Demography 34, 31-48, and references therein*

---

# Population Models. Conclusion

---

- The collection of Population Models for modeling and analyzing the health effects are developed:
    - *Generalized frailty models, correlated frailty models*
    - *Quadratic Hazard Model*
    - *Stochastic Process Model*
    - *Microsimulation models*
    - *Models for continuous tracking data analyses*
    - *Linear Latent Structure Models*
  - The procedures presented have been successful and should be effectively migrated into this new application area.
  - These models are useful and directly applicable to statistical examination of populations exposed to ionizing radiation.
  - In progress: simulation experiments where data with structure of databases on exposed populations are simulated and the models are applied to extract parameters used for the generation of the data.
-

---

*Part II:*

***Bio-Molecular Models***

# General Concept

---

The current state of the art of modeling efforts in tumorigenesis relies on a multi-stage hypothesis that is implemented in multi-stage models of carcinogenesis (UNSCEAR, 2001).

Three stages of the process of tumor development:

1. formation of initiated cells;
2. promotion and neoplastic conversion of initiated cells resulting in appearance of the first malignant clonogenic cell;
3. subsequent growth and progression of a malignant tumor.

The duration of each stage of carcinogenesis is thought of as a random variable.

The underlying idea of these models is the concept of sequential, interacting gene mutations as the driving force of tumorigenesis.

# Mathematical Formalism

---

Principal Aim is to predict a hazard function (e.g., incidence, mortality, survival, etc) in terms of biological parameters.

Typical sequence of steps of mathematical modeling

- 1. Define cell processes to be taken into account (e.g., specify compartment structure, make assumptions on proliferation, apoptosis, mutation rates);*
- 2. Define whether each process occurs stochastically or deterministically;*
- 3. Write corresponding differential equations (DE), e.g. i) ordinary DE for the numbers of cells in a compartment for deterministic models, or ii) stochastic DE for numbers of cell or partial (Kolmogorov) DE equations for probability generating functions for stochastic model;*
- 4. Solve these equations and use the solution (in the form of deterministic function, or in the form of stochastic process, or in the form of p.g.f.) to predict a hazard rate.*

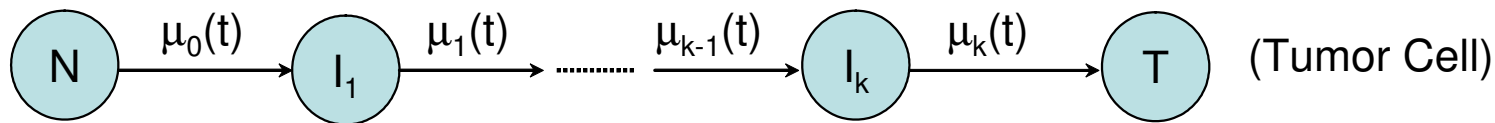
# Mechanistic Carcinogenesis Models

---

- Multi-Stage models of Nordling (1953) and Armitage-Doll (1954)
- Two-mutation model of Moolgavkar, Venzon and Knudson (1979, 1981)
- Two stage clonal expansion model (TSCE)
- Generalized Multistage Models
- Models based on ideas of queuing theory
- Stochastic models developed by Tan and colleagues (1991-2005)
- Multiple Pathways models
- Recent generalizations

# Nordling and Armitage-Doll models

Aim: to explain the observation that age-specific rates of many common carcinomas increased roughly with a power of age. They assumed that cancer is the result of the accumulation of a critical number of mutations. The order of mutation can be important (Armitage-Doll) or not (Nordling).



In both approaches the incidence function is approximately equal

$$I(t) = ct^k$$

$$c = N\mu_0\mu_1 \dots \mu_k/k! \quad \text{for Armitage-Doll model}$$

$$c = (k + 1)N\mu_0\mu_1 \dots \mu_k \quad \text{for Nordling model}$$

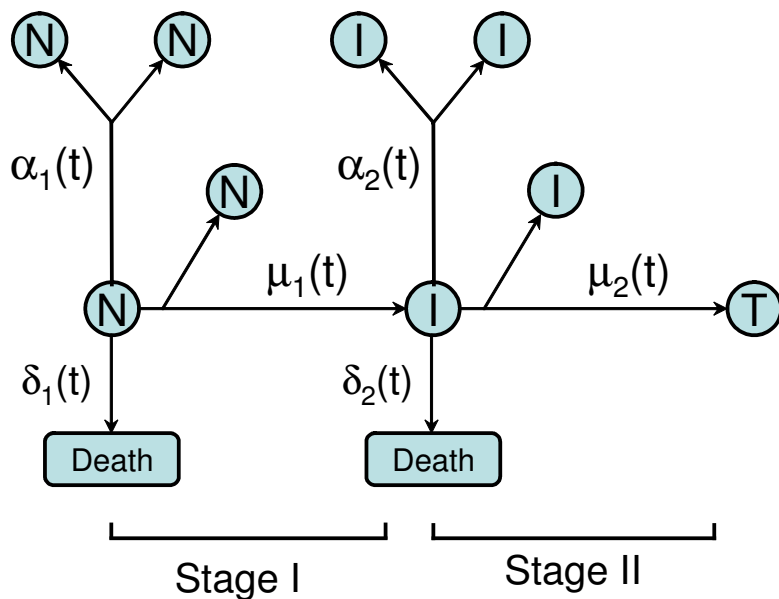
$\mu$ 's are time independent and  $N$  is the total number of susceptible cells.

# Two-Stage Models: Moolgavkar-Venzon-Knudson model

Armitage-Doll model predicts 4-7 stages, however there is a little evidence that there are as many stages as this (BEIR V).

Aim: to reduce the biologically implausible number of stages.

A series of generalizations (Armitage, Doll 1957; Knudson 1971; Moolgavkar, Venzon 1979) result in a two-stage model known as Moolgavkar-Venzon-Knudson model.



$N$  — normal (stem) cells;

$I$  — intermediate cells;

$T$  — malignant cells;

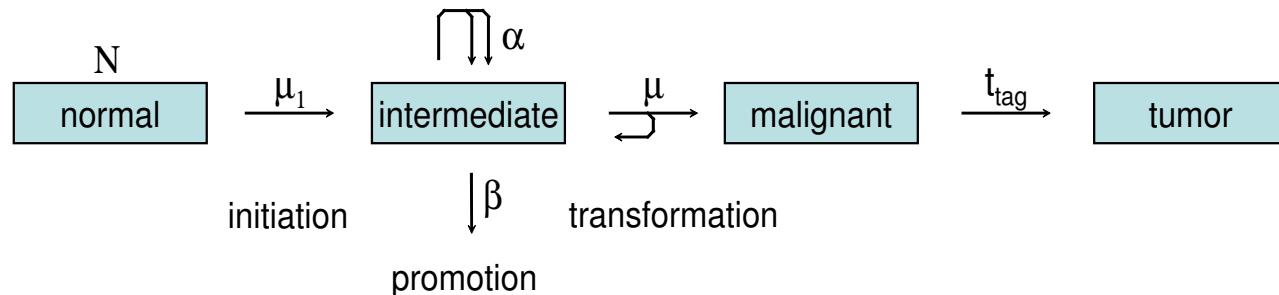
**Stage I** describes initiation

**Stage II** describes  
promotion and conversion

# Two Stage Clonal Expansion model

The most popular version of the two-stage model is the Two Stage Clonal Expansion (TSCE) model which additionally assumes

- Number of normal cells is either constant or described by a deterministic function.
- All rates are time independent.



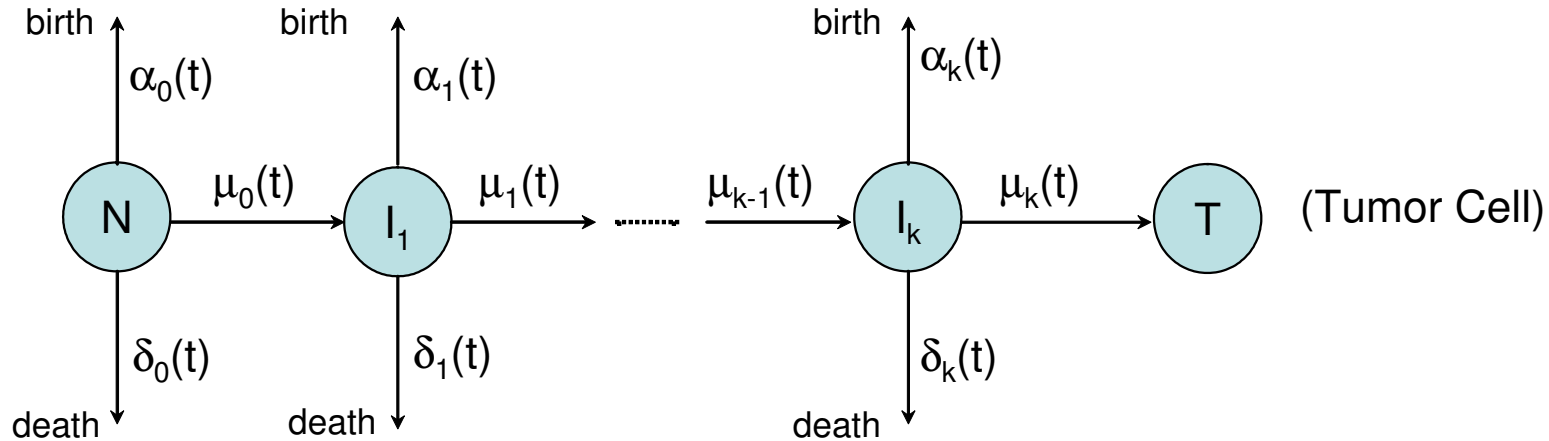
Attractive property: *Spontaneous hazard rate can be expressed analytically in terms of only three parameters (Heidenreich, Paretzke, Rad.Res 156, 678-681 (2001)):*

$$h(t) = \frac{X(e^{(\gamma+2q)t} - 1)}{q(e^{(\gamma+2q)t} + 1)\gamma}; \quad X = N\mu\mu_1, \quad \gamma = \alpha - \beta - \mu, \quad q = \frac{1}{2}(-\gamma + \sqrt{\gamma^2 + 4\alpha\mu}).$$

Main disadvantage: *Not all biological parameters (i.e.,  $N$ ,  $\mu$ ,  $\mu_1$ ,  $\alpha$ ,  $\beta$ ) can be identified using the data on age-specific incidence rates.*

# Multistage models

Armitage-Doll and two-stage models can be generalized to allow for an arbitrary number of mutation stages (Little, 1995).



Analytical solution is no longer possible

$$h(t) = - \int_0^t \mu_0(u) N(u) \frac{\partial \Phi(t, u)}{\partial t} du$$

where  $\Phi(t, u)$  has meaning of a p.g.f. and satisfies the Kolmogorov backward equation. This equation is solved numerically.

# Stochastic models of Klebanov, Yakovlev with colleagues

---

Klebanov, Rachev, and Yakovlev (1993) developed a model for radiation carcinogenesis based on assumptions:

- *Number of lesions formed by IR (accumulated by time  $T$ ) is the Poisson variable with expectation  $\gamma T$ , where  $\gamma$  is related to IR dose.*
- *Lesions are subject to repair process. Repair system is modeled as  $M/M/m$  queue with losses. Then probability for a lesion not to be served is modeled as a well-known probability of “losing a customer.”*
- *Each promoted lesion ultimately gives rise to an overt tumor after a certain period of time.*

Yakovlev and Polig (1996) generalized this model by an incorporation of radiation induced cell death. Hazard is quite complex function of dose rate and p.d.f. of promotion time distribution. This model, when promotion time is modeled by  $\gamma$ -distribution, and TSCE were tested with several sets of epidemiologic data (Gregori et.al. 2002). No conclusion about the better model was made.

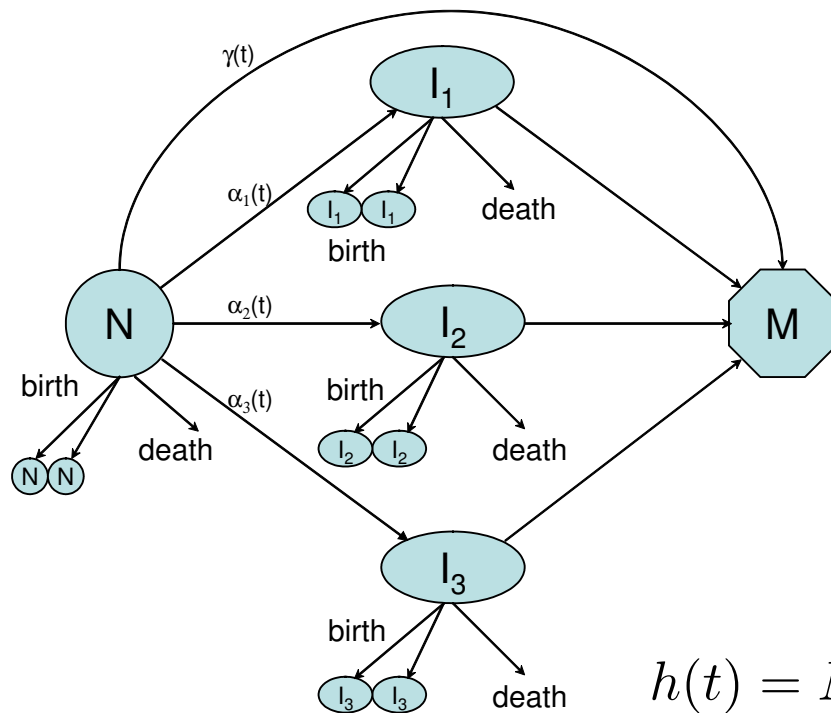
# Stochastic models of Tan and colleagues

---

- *Tan (1991) has developed a number of generalizations of two-mutation models of MVK and has documented biological evidence and mathematical formalism.*
- *Since most of these models are complicated far beyond the scope of the MVK two-stage model, Tan and Chen (1998) developed state space models (Kalman filter models) for carcinogenesis using formalism of stochastic differential equations. They also demonstrated how their formalism is related to classical formalism based on p.g.f.*
- *Tan, Zhang, and Chen (2004) developed advanced statistical procedures to estimate the unknown parameters of the state space model via multi-level Gibbs sampling method (i.e., using MCMC) and applied these procedures to the British physician data on lung cancer due to smoking.*
- *Tan (2005) has shown that under some mild conditions, the number of initiated cells in an extended two-stage model of carcinogenesis can be approximated by a diffusion process.*

# Multiple Pathways models

In many cases, it was observed that the same cancer may arise from different carcinogenic processes. Careful examination of the literature of cancer biology would reveal that multiple pathways for cancer may be quite common in real life (Tan, 1991). More recent biological evidence were documented by Tan and Chen (1998).



Roughly, hazard rate is the sum of possible pathways.

For simplest paths the hazard rate can be calculated analytically.

$$h(t) = N(t)\gamma(t) + \sum_{i=1}^3 \int_0^t N(u)\alpha_i(u) \frac{\partial \phi_i(t, u)}{\partial t} du$$

# Description of IR induced carcinogenesis

---

- *IR can induce specific mutations in stem cells and therefore increase the number of intermediate cells susceptible to further stages of carcinogenesis*
- *IR can also have promoting action to carcinogenesis. Basic argument is that inactivated by IR, stem cells may be replaced by the division of stem cells in which intermediate cells have a growth advantage (Heidenreich et.al., Rad. Res. 2001, 155, 870-872).*

A typical way to incorporate these effects into mechanistic models is to assume that rates of initiation, promotion, and conversion become dose-dependent.

Recently, such effects were analyzed and discussed for radon-induced lung cancer in Colorado Plateau uranium miners (Little et al., 2002) and French and Czech miner cohorts (Brugmans et al., 2004; Heidenreich et al., 2004).

Further discussions (*Bijwaard et al. 2005; Heidenreich 2005, Laurier et al. 2005*) covered several aspects: *biological viability of the models, testing hypotheses about the processes of radiation carcinogenesis, selection of the best fitting model, comparison to the empirical approach which uses statistical modeling in describing the data, etc.*

Conclusion of the discussion: *Even if biologically-motivated mathematical models of carcinogenesis are necessarily a crude simplification of the biological reality, such models constitute a complementary approach to empirical statistical models.*

# Uncertainties of TSCE and possible generalizations

---

TSCE being one of the most popular models of IR induced carcinogenesis has limitations

- *Parameter identifiability. Only three combinations of biological parameters are identified from the age-specific hazard function.*
- *Oversimplified biological mechanisms.*
- *Usage of parameters which cannot be directly measured.*
- *Not clear how to predict individualized risks.*

One possible solution is to combine data on the age-specific hazard function with additional measurements indirectly related to the model parameters, e.g., to measure apoptosis rate. This is especially important in the light of recent results of Akleyev et al (2006) demonstrated strong dose dependence of apoptosis rate for individuals from the ETRC.

# Generalizations: Barrier Mechanisms

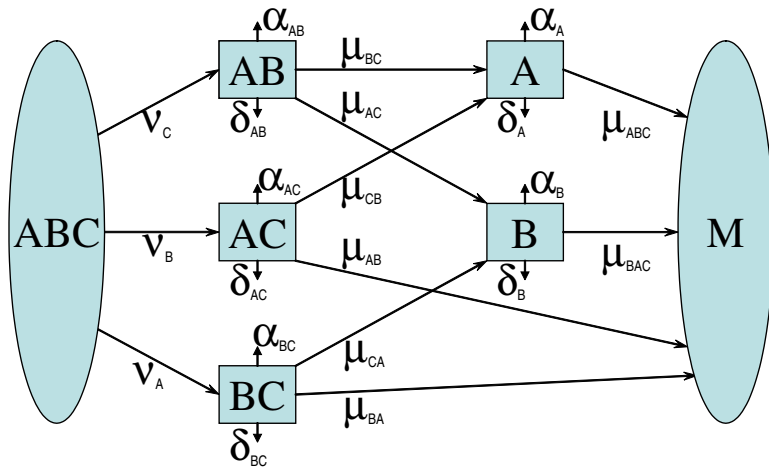
---

Concept of barrier mechanisms is based on the following facts/observations/principles:

- ➔ *Basis of the development of the carcinogenesis is a set of structure-functional changes in radiosensitive cells ultimately expressed at the higher level of hierarchy of the human body.*
  - ➔ *Effects at the cell level are caused by the quality of operation of complex of barrier systems (antioxidant defense, repair systems, apoptosis, and others), which are in complex interaction with each other.*
  - ➔ *Cell malignization occurs due to inefficient operation of a part of, or all, barrier systems.*
  - ➔ *Efficient operation results in the development of adaptation at cell level.*
  - ➔ *Hierarchy and complex interaction of barrier systems can compensate the inefficiency in operation of certain systems by reinforcing others, and as a result decrease the risk of pathology development.*
-

# Barrier Mechanisms: Carcinogenesis Model

Our new model is a generalization of ideas of multistage and multiple pathways models on the concept of barrier mechanisms



*Transfers correspond to the failure of a specific barrier, i.e., apoptosis (A), repair (B), antioxidant defence (C).*

*Two types of measures are expressed in terms of model parameters: age-specific hazard rate and age-specific means of state of each barrier*

$$dX_N(t) = -(\nu_A + \nu_B + \nu_C)X_N(t) + b_N dw_t$$

$$dX_{BC}(t) = (\alpha_{BC} - \delta_{BC})X_{BC}(t) + \nu_A X_N(t) + b_{BC} dw_t$$

$$dX_{AC}(t) = (\alpha_{AC} - \delta_{AC})X_{AC}(t) + \nu_B X_N(t) + b_{AC} dw_t$$

$$dX_{AB}(t) = (\alpha_{AB} - \delta_{AB})X_{AB}(t) + \nu_C X_N(t) + b_{AB} dw_t$$

$$dX_B(t) = (\alpha_B - \delta_B)X_B(t) + \mu_{AC} X_{AB}(t) + \mu_{CA} X_{BC}(t) + b_B dw_t$$

$$dX_A(t) = (\alpha_A - \delta_A)X_A(t) + \mu_{CB} X_{AC}(t) + \mu_{BC} X_{AB}(t) + b_A dw_t$$

$$dX_M(t) = \mu_{AB} X_{AC}(t) + \mu_{BA} X_{BC}(t) + \mu_{ABC} X_A(t) + \mu_{BAC} X_B(t) + b_M dw_t$$

*Solution of this system is normal (i.e., diffusion) stochastic process.*

# Barrier Mechanisms: Future Perspectives

---

The following Aims are in progress or will be addressed in future projects

- *Perform detailed simulation studies to develop the design of future projects.*
- *Generalize this model to include dose-dependent rates and apply it to ETRC and other sets of epidemiologic data.*
- *Apply this model to combined dataset of epidemiologic data and biomolecular measurements associated with barrier mechanisms available at the URCRM.*
- *Based on the results, develop an individualized model of development of IR induced leukemia and solid cancers.*
- *For leukemia model, include the model of different stages of differentiation of blood cells: i.e., describe the process from red-bone stem cells to cells of peripheral blood.*
- *Include other potentially important barriers, e.g., telomerase, growth arrest, etc.*
- *Use the concept to develop the model of IR induced genomic instability.*
- *Generalize this model to include a description of non-cancer and deterministic effects (ICRP publications 41, 58).*

# Possible Application: Biodosimetry

---

There are no ideal biological markers of protracted IR dose.

Reasons:

- *Presence of intracellular barrier systems repairing genetic damages or eliminating cells with pathologic changes;*
- *Bystander effects and IR induced genomic instability can distort the dose-effect dependence.*

Further ways of development of biodosimetry for protracted irradiation:

- *Further search markers at molecular and cellular levels minimally subjected to intergroup fluctuations;*
- *Modeling dynamics of onset, development, and conservation of cell pathology (i.e., a dose marker) taking into account the concept of barrier systems;*
- *Possible extraction of stable effect using specific statistical technology (e.g., latent structure analyses) from high-dimensional data (e.g., gene expression data).*

# Conclusion

---

- *Biologically-motivated mathematical models of carcinogenesis constitute a complementary approach to empirical statistical models, and therefore, are important in understanding the carcinogenesis process in spite of simplification of the biological reality.*
- *Broadly used mechanistic models have certain disadvantages, e.g., oversimplification of two-stage modeling, difficult mathematical formalism of exact multistage models, and only partial identifiability of biological parameters in many of them.*
- *Recent developments of Tan and colleagues provided new techniques for describing multistage and multiple pathways models and provided background for constructing models in the form of diffusion stochastic processes.*
- *The concept of barrier mechanism and the related models of carcinogenesis developing in Duke in collaboration with the URCRM have the potential to overcome these difficulties and perspectives in construction of individualized models of cancer and non-cancer health effects induced by IR.*

---

*Part I:*

**Appendix: Population Models**  
**Full Version**

# Databases used for the constructing population models

---

- Framingham Heart Study, 50-year follow-up, *hundreds of risk factors measured biannually; 5209 persons at the inception*
- National Long Term Care Survey (1982, 1984, 1989, 1994, 1999, 2004); *41,947 different individuals, with roughly 70,000 screening interviews and 28,500 detailed interviews; 26,000 deaths recorded by 2003; about 400,000 person years*
- Medicare Service Use Files from 1982 to 2001. *Continuous Medicare history files contain information about costs, treatments and diagnoses on the dates delivered, and mortality data until the date of death.*
- Multiple Cause of death: 1968-2001; *65 mln death records included underlying and secondary causes of death*
- The Surveillance, Epidemiology and End Results (SEER) (*began in 1973 and captures approximately 14% of the US population*), *detailed clinical information collected about each incident cancer diagnosis*

# Classical Population Models

---

- Gompertz (1825)  $\mu_0(t) = \mu_G(t) = \alpha \exp \theta t$
- Weibull,  $\mu_0(t) = \mu_W(t) = \alpha t^{m-1}$
- Strehler and Mildvan model (1960)  $\mu = K \exp(V(x)/\epsilon D)$ ; vitality  
 $V(x) = V_0(1 - Bx)$
- Sacher and Trucco model, (1962) – stochastic mechanism of aging and mortality
- Frailty models; correlated frailty models; debilitation models; repair capacity models (reviewed by Yashin et al, 2000)

# Modern Population Models

---

- Extensions of frailty models
- Event history models
- Quadratic hazard models
- Stochastic process models
- Microsimulation models
- Models of latent structure analyses

## Extension of frailty models

---

Vaupel, Manton, Stallard (1979) *Demography*. 16: 439-54

Manton, Stallard, Vaupel (1986) *JASA* 81: 635-644.

Manton, Lowrimore, Yashin (1993) *Math.Pop.St.* 4:133-47.

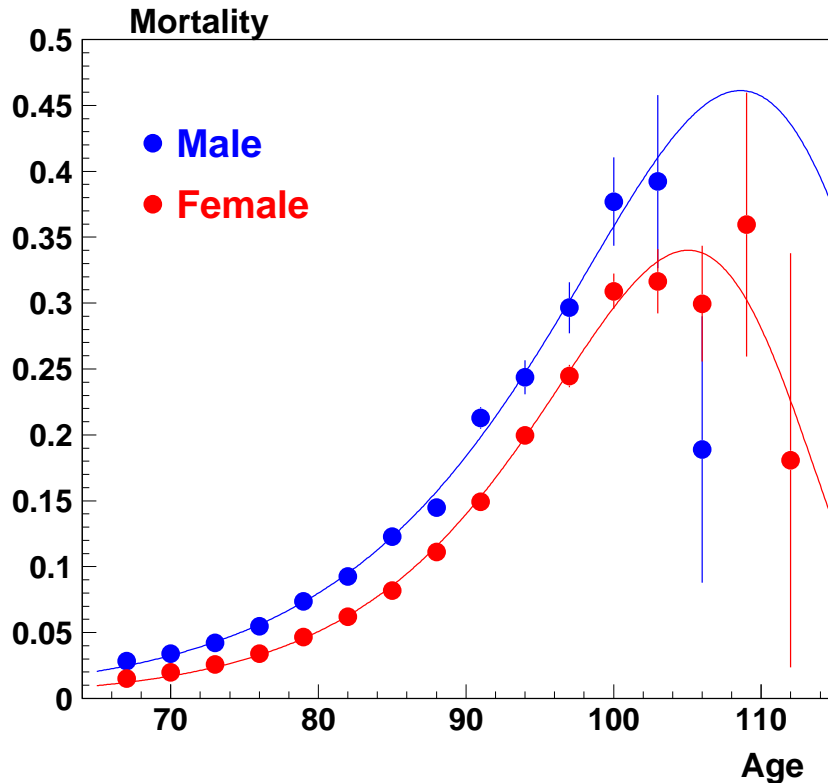
$$\mu(t) = \frac{\mu_0(t)}{\left[1 + n\gamma \int_0^t du \mu_0(u)\right]^{1/n}}; \quad \mu_0(t) = \mu_G(t) = \alpha \exp \theta t$$
$$\mu_0(t) = \mu_W(t) = \alpha t^{m-1}$$

$n = 1$  –  $\gamma$ -distribution

$n = 2$  – inverse Gaussian distribution

# Model: Fitting

Best fit is obtained for  $n = 0$



Questions:

1. Can we have explicit form for  $\mu(t)$ ?
2. Does distribution corresponding to  $n = 0$  exist?

## Model: Explicit results

---

Generalized Weibull model:

$$\mu_W(t, t_0, m, \gamma) = \frac{m-1}{t_0 \gamma} \exp\left(-\left(\frac{t}{t_0}\right)^m \frac{m-1}{m}\right) \left(\frac{t}{t_0}\right)^{m-1}$$

Generalized Gompertz model:

$$\mu_G(t, t_0, \theta, \gamma) = \frac{b\theta}{\gamma} \exp(b(1 - e^{\theta t}) + \theta t) \quad b = e^{-\theta t_0} \frac{\theta}{e^\theta - 1}$$

# Quadratic hazard model

---

- Structure of data: follow-up with measurements of risk factors at regular time intervals
- Mortality is quadratic function of covariates
- Time to failure (death) of an organism is modeled by a random walk model with "manholes". The stochastic differential equations for the random walk are

$$dx_w(t) = u(x_w, t)dt + d\xi(x_w, t),$$

$$dP(x_w) = \mu(x_w, t)P(x_w)dt$$

This equation represents changes for organism  $w$  on each of  $n$  coordinate dimensions,  $x = (x, j = 1, 2, \dots, n)$  at time  $t$ .  $P$  is probability to survive for  $w$ .

---

# Quadratic hazard model

- Kolmogorov-Fokker-Planck equation

$$\frac{\partial f}{\partial t} = - \sum_j u_j \frac{\partial f}{\partial x_j} - f \sum_j \frac{\partial u_j}{\partial x_j} + \frac{1}{2} \sum_i \sum_j \sigma_{ij}^0 \frac{\partial^2 f}{\partial x_j \partial x_j} - \mu f$$

- Mortality is a quadratic function of risk factors

$$\mu = \mu_0 + b_t x + \frac{1}{2} x^T B_t x$$

- Mean  $\nu$  and covariance matrix  $V$ . Total/survival:

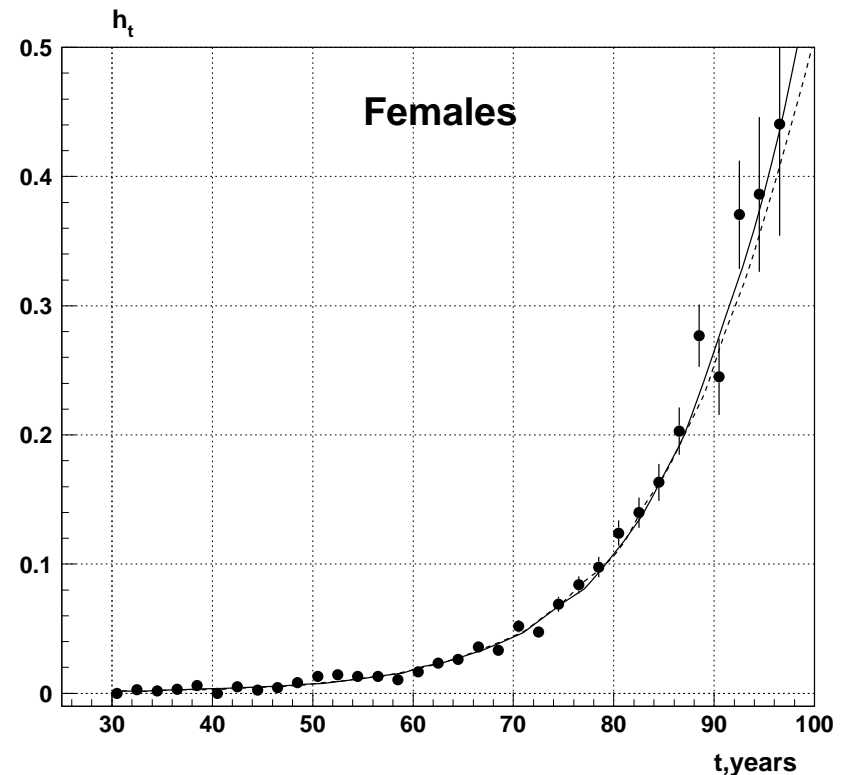
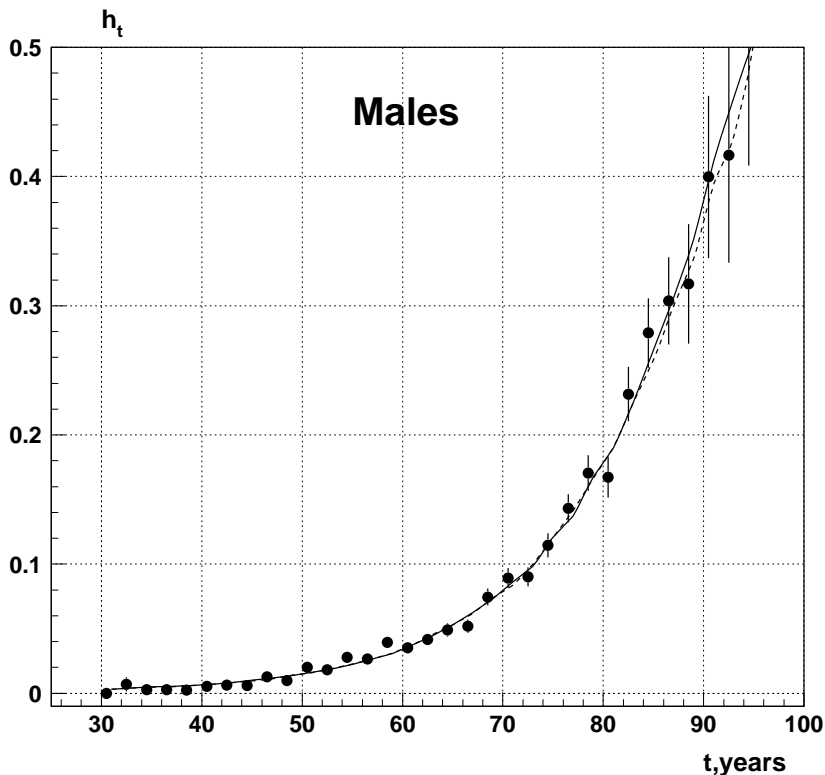
$$\nu_t^* = \nu_t - V_t^* (b_t + B_t \nu_t), \quad V_t^* = (V_t^{-1} + B_t)^{-1}.$$

- dynamics  $\nu_{t+1} = u_0 + R\nu_t^* \quad V_{t+1} = \Sigma + R V_t^* R^T$

- Likelihood  $L = \prod_i p(X^i, T^i)$

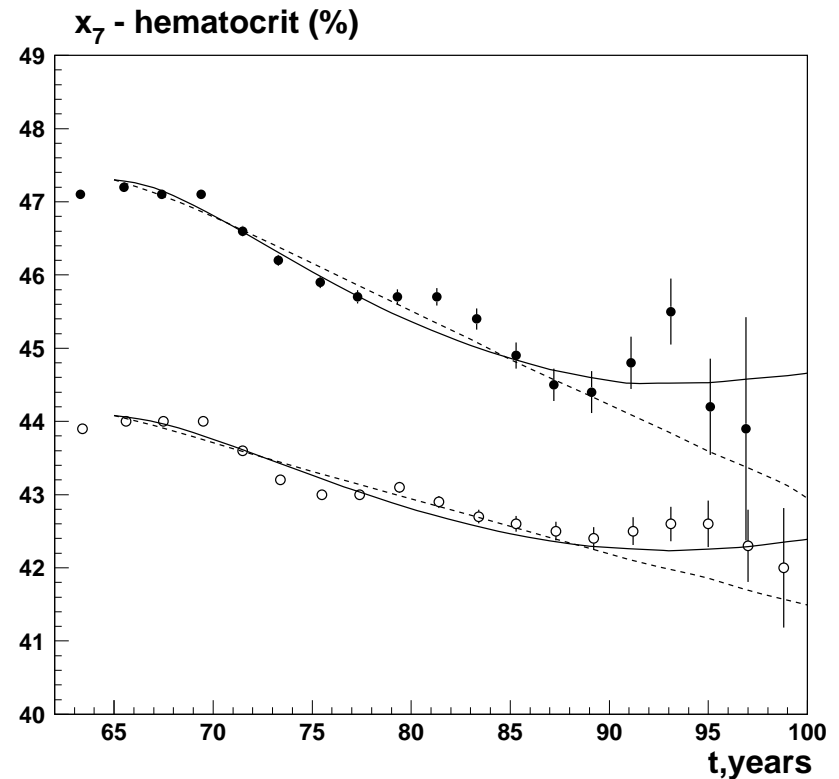
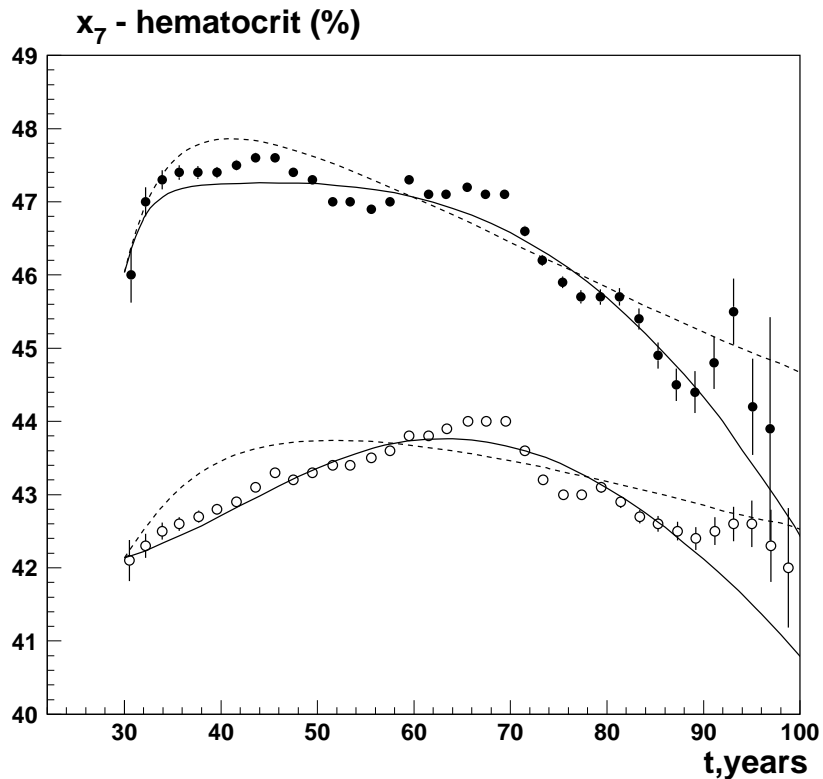
$$p(X, T) = p_1(\mathbf{x}_{t_0}) \left\{ \prod_{t=t_0}^{T-1} S(t|\mathbf{x}_t) \phi(\mathbf{x}_{t+1}|\mathbf{x}_t) \right\} \{1 - S(T|\mathbf{x}_T)\}$$

# Quadratic hazard model: mortality prediction



Probability of death within two years estimated from Framingham data (filled dots with error bars) for males (left) and females (right), with theoretical predictions given by linear (dashed line) and nonlinear (solid line) models

# Quadratic hazard model: dynamics



Projections from 30 and from 65 years for hematocrit for males (filled dots, upper curves) and females (open dots, lower curves).

# Stochastic process model: Advantages

---

Alternative approach to make these projection were described in paper by Yashin and Manton, *Statistical Science* 1997, 12, 20

- All parameters are defined within jointed likelihood function, so only one numerical procedure is needed to estimate all parameters simultaneously.
- It is not necessary to attract any procedures to fill missing data
- The projection is obtained as solution of differential equation, so time between surveys can be not fixed. Therefore, it is possible to make predictions within 'year by year' or 'month by month' scheme.

# Stochastic process model

---

- ➔ mortality is a quadratic function of  $x(t)$

$$\mu(x(t), t) = \mu_0(t) + 2b(t)x(t) + x^*(t)B(t)x(t)$$

$$dx(t) = (a_0(t) + a_1(t)x(t))dt + a_2(t)dW_t$$

- ➔ Likelihood:

$$L = \prod_{i=1}^N \hat{\mu}(\tau_i, \hat{x}(\tau_i))^{\delta_i} \exp \left( - \int_0^{\tau_i} du \hat{\mu}(u, \hat{x}_i(u)) \right) \times \prod_{j=1}^{k_i} f(x_i(t_j) | \hat{x}_i(t_{j-1}));$$

where  $\hat{\mu}(\hat{x}(t), t) = m^*(t)B(t)m(t) + 2b(t)m(t) + \text{tr}(B(t)\gamma(t)) + \mu_0(t)$ ,

- ➔ System of differential equations:

$$dm(t)/dt = a_0(t) + (a_1(t) - 2b(t))m(t) - 2\gamma(t)B(t)m(t)$$

$$d\gamma(t)/dt = a_1(t)\gamma(t) + \gamma(t)a_1^*(t) + a_2(t)a_2^*(t) - 2\gamma(t)B(t)\gamma(t)$$

# Microsimulation Models

---

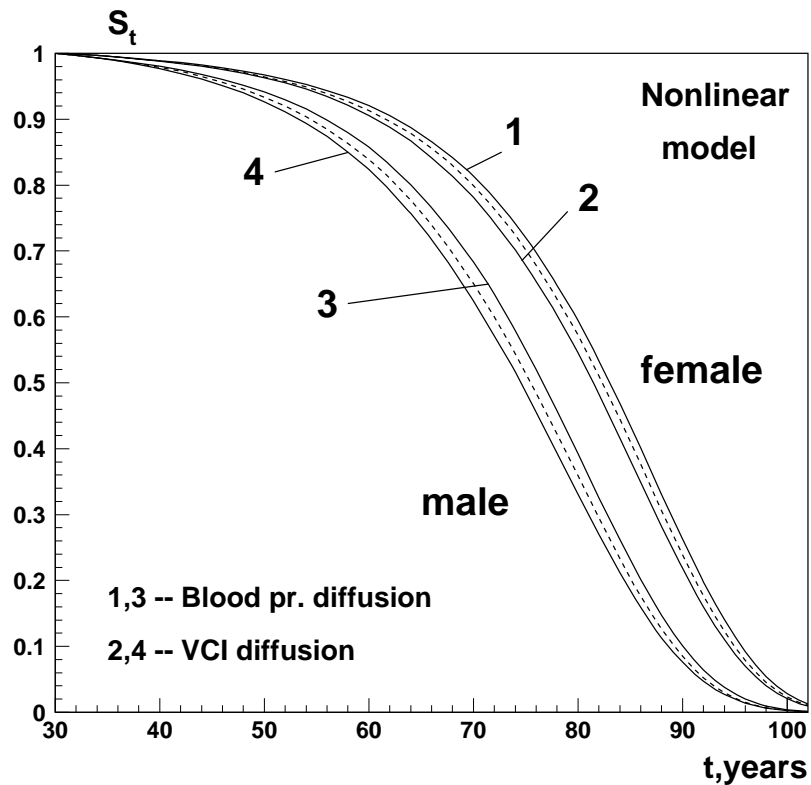
## Basic advantages

- ➔ work with complicated dynamic model (e.g., nonlinear, history)
- ➔ investigate medical interventions

## Basic steps:

- ➔ simulate theoretical population (i.e. 100,000 individuals with normally distributed risk factors)
- ➔ calculate mortality and survival function for each individuals at the each time interval; define randomly if this individual survives in this time interval or not
- ➔ simulate risk factors for survival individuals for the next age

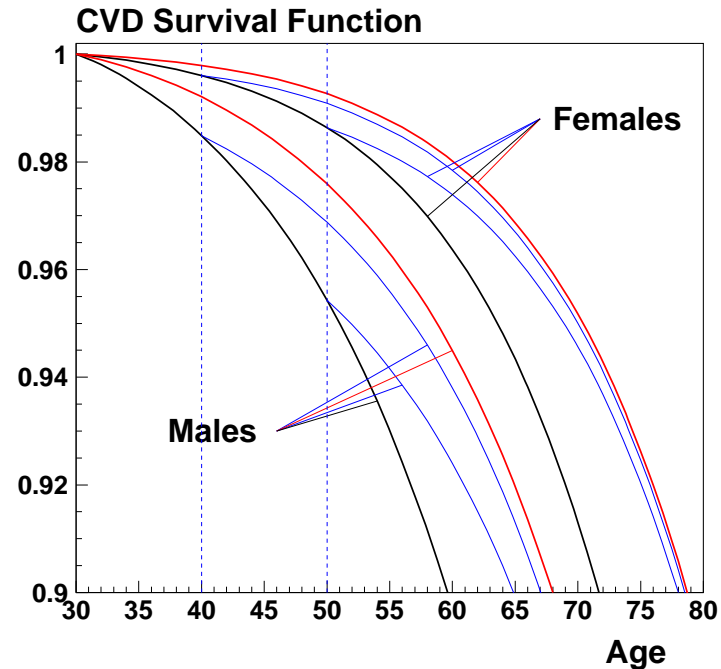
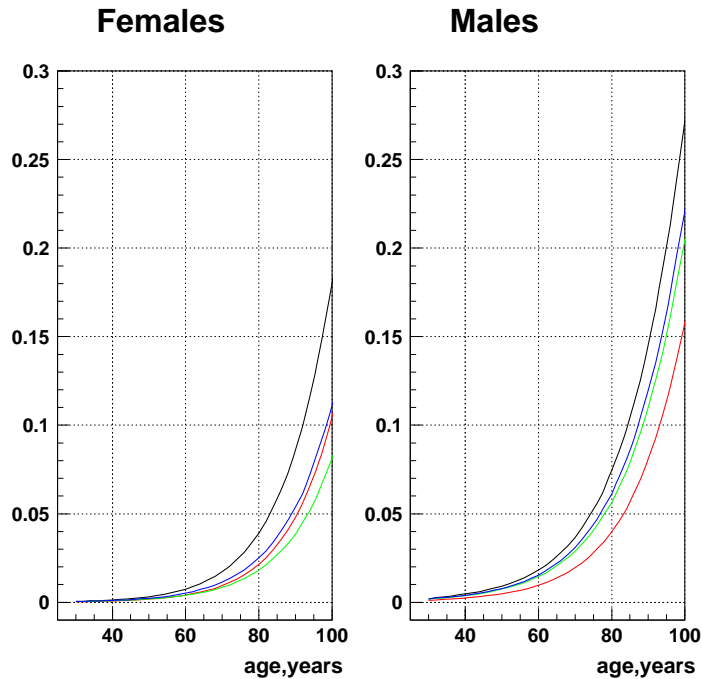
# Microsimulation models



- ➔ diffusion for VCI is decreased by a factor of 2
- ➔ blood pressure control, i.e. the diffusion vectors for pulse pressure and diastolic blood pressure are lowered by a factor of 2.

# Microsimulation Models: Interventions

Kravchenko et al., SAGE KE 2005 Jun 22; 2005 (25):pe18

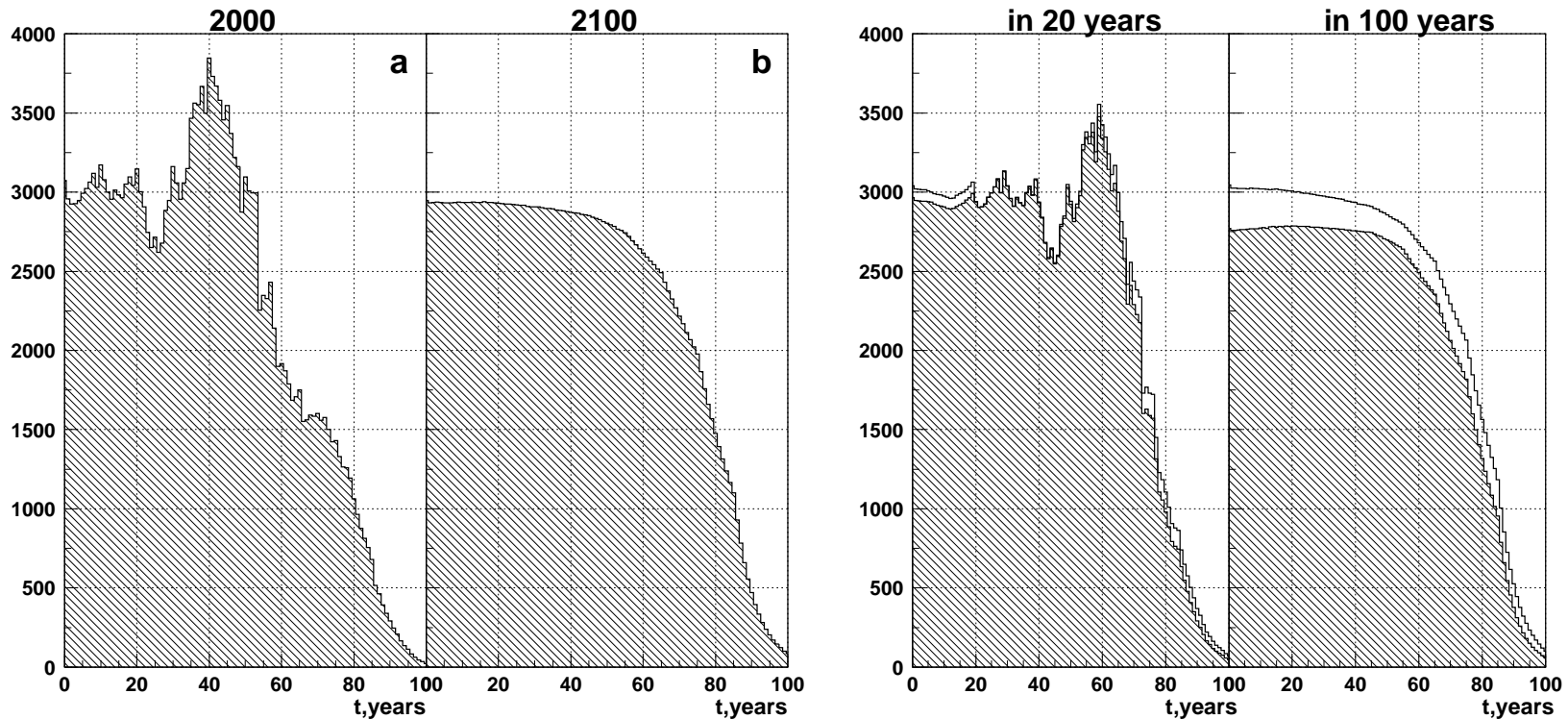


CVD mortality for four scenarios: without interventions (black), progenitor cell therapy (red), reduction (blue), and "ideal" restriction (green) of CVD risk factors

CVD survival function without interventions (black) and for progenitor cell therapy at age 30 (red) and 40 and 50 (blue) years old

# Microsimulation Models: Interventions—smoking level

Smoking level → infertility and mortality rates



Age distribution of the white American population with current smoking prevalence, years 2000 and 2100 (left) and in 20 and in 100 years, according to the smoking prevalence: shaded areas – doubled prevalence of smoking, transparent – 50% reduced prevalence of smoking in the population (right).

# LLS analysis: the problem

---

Given:

- ➔ Measurements with discrete outcome made on individuals

Find:

- ➔ Properties of the population
- ➔ Properties of the individuals

*It happens that both goals may be achieved only simultaneously, and to increase precision one has to increase both sample size and number of measurements.*

# LLS analysis: input data

Results of  $J$  measurements made on  $N$  individuals:

Individual 1	$x_1^1$	...	$x_J^1$
...		...	
Individual $i$	...	$x_j^i$	...
...		...	
Individual $N$	$x_1^N$	...	$x_J^N$

Outcomes of the 1<sup>st</sup> measurement

Outcomes of the  $J^{\text{th}}$  measurement

For every  $i$ , possible outcomes of  $j^{\text{th}}$  measurement are:

$$1, \dots, L_j$$

## LLS analysis: Mathematics

---

Random variables  $X_1, \dots, X_J$ ;  $X_j$  takes values in  $\{1, \dots, L_j\}$ .

The joint distribution is given by elementary probabilities:

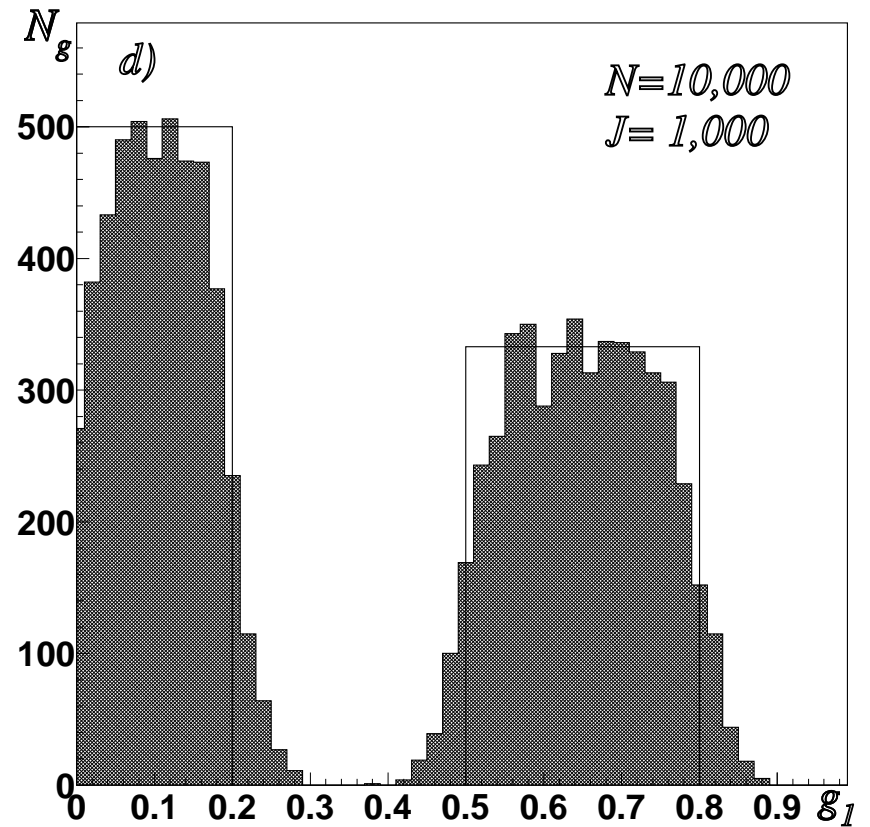
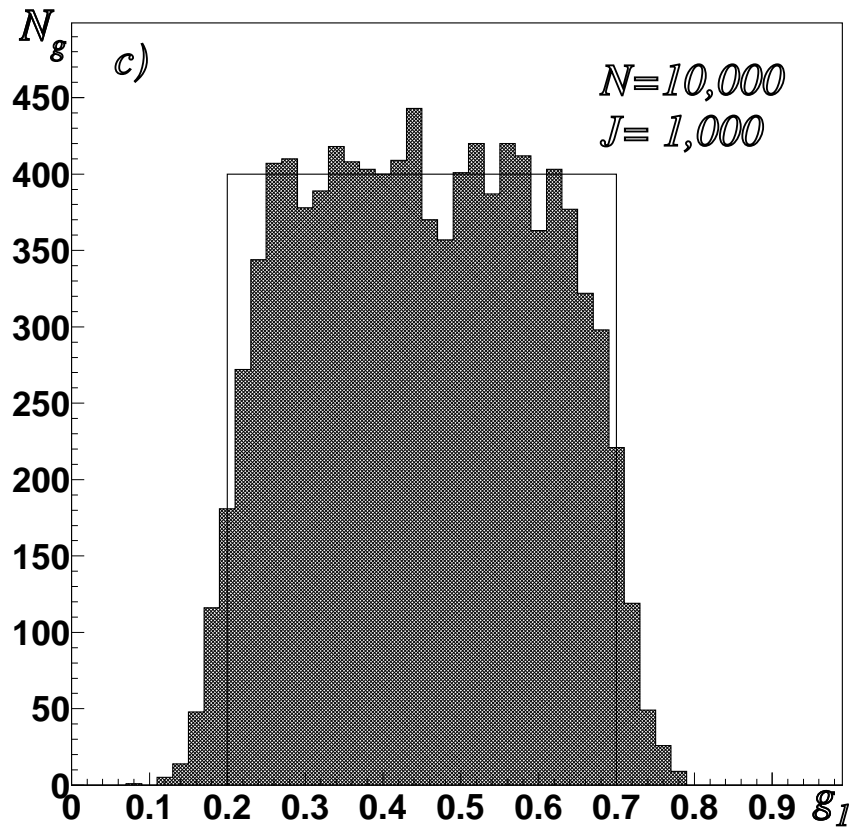
$$p_{\ell} = \Pr(X_1 = \ell_1 \text{ and } \dots \text{ and } X_J = \ell_J)$$

$\ell = (\ell_1, \dots, \ell_J)$  — response pattern

*These and only these values are directly estimable from the observations. Frequencies  $f_{\ell} = \frac{N_{\ell}}{N}$  are consistent estimators for  $p_{\ell}$ .*

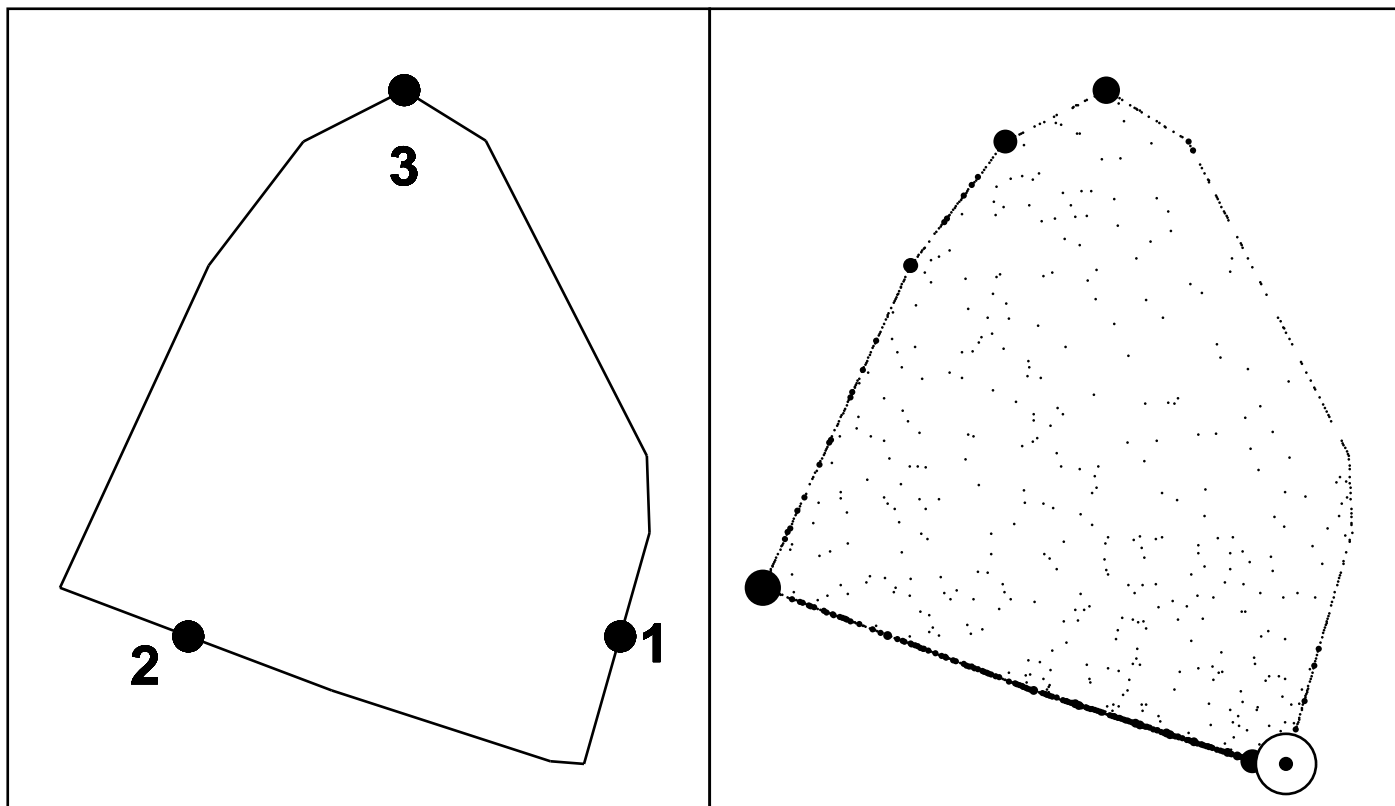
*In general, the joint distribution cannot be described by smaller number of parameters.*

# Simulation studies: case of continuous distribution



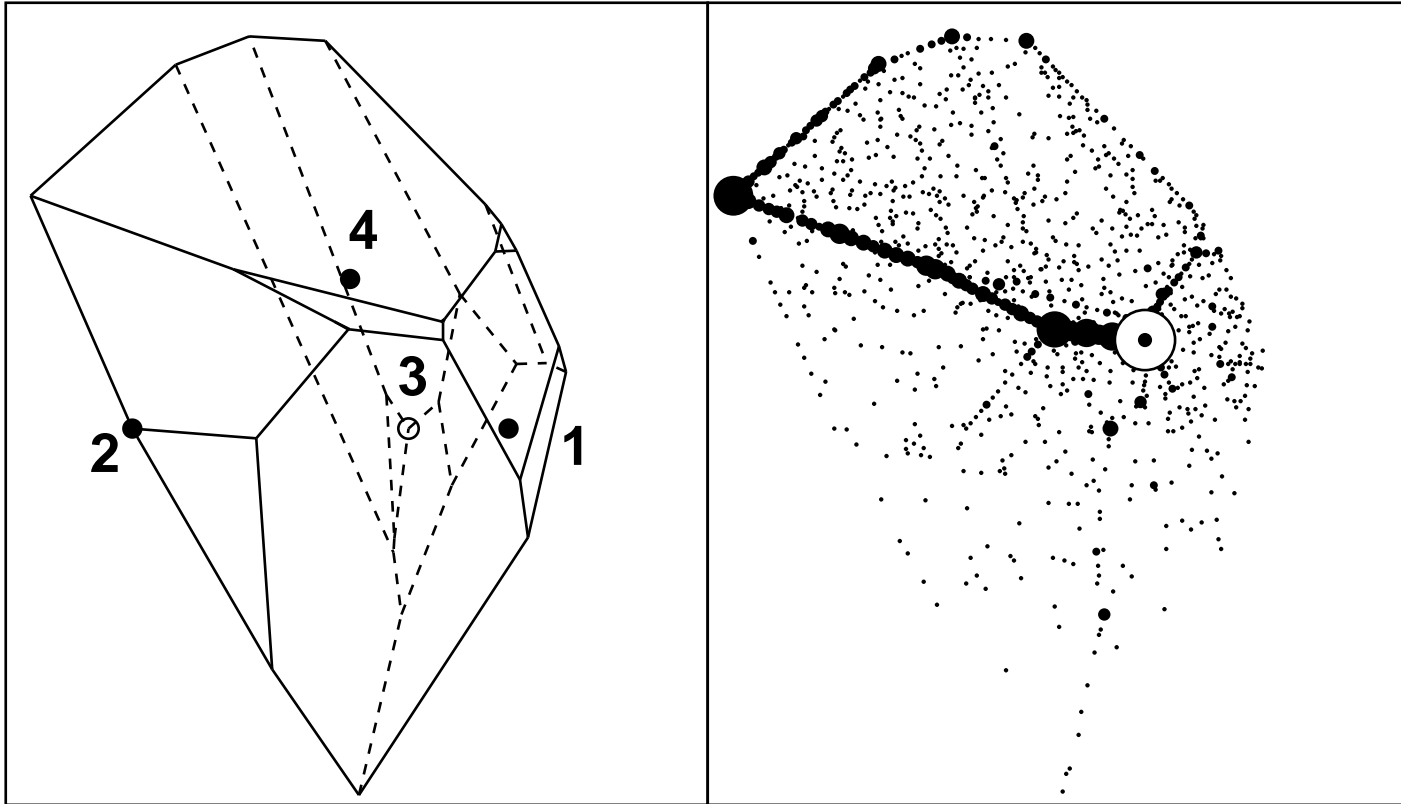
The mixing distribution is shown by a solid line.

# LLS Results: 3-dimensional heterogeneity



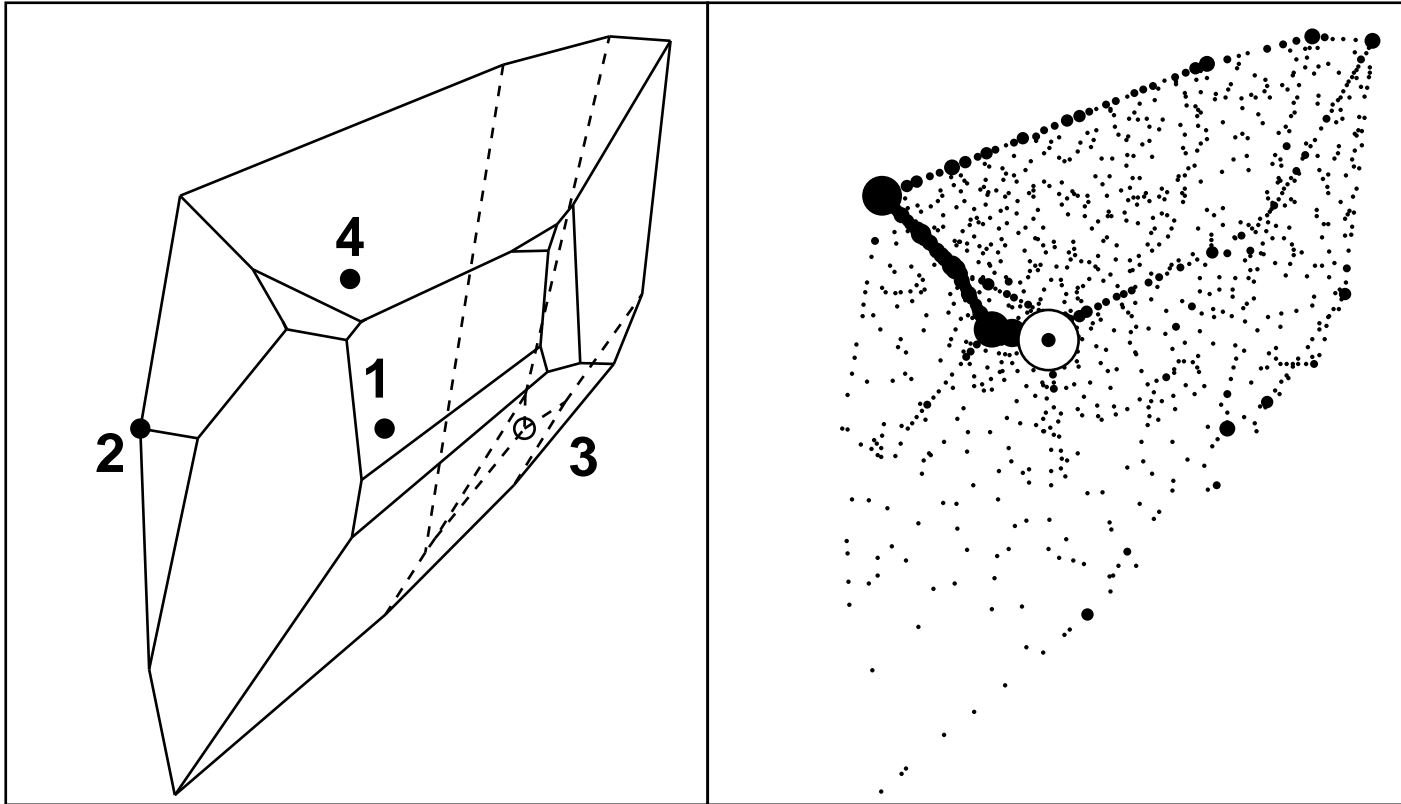
- 1—healthy
- 2—disabled
- 3—with chronic diseases

# LLS Results: 4-dimensional heterogeneity



- 1—healthy
- 2—disabled
- 3—with chronic diseases
- 4—partly disabled

# LLS Results: 4-dimensional heterogeneity



- 1 — healthy
- 2 — disabled
- 3 — with chronic diseases
- 4 — partly disabled

## Methods: disease onset identification

---

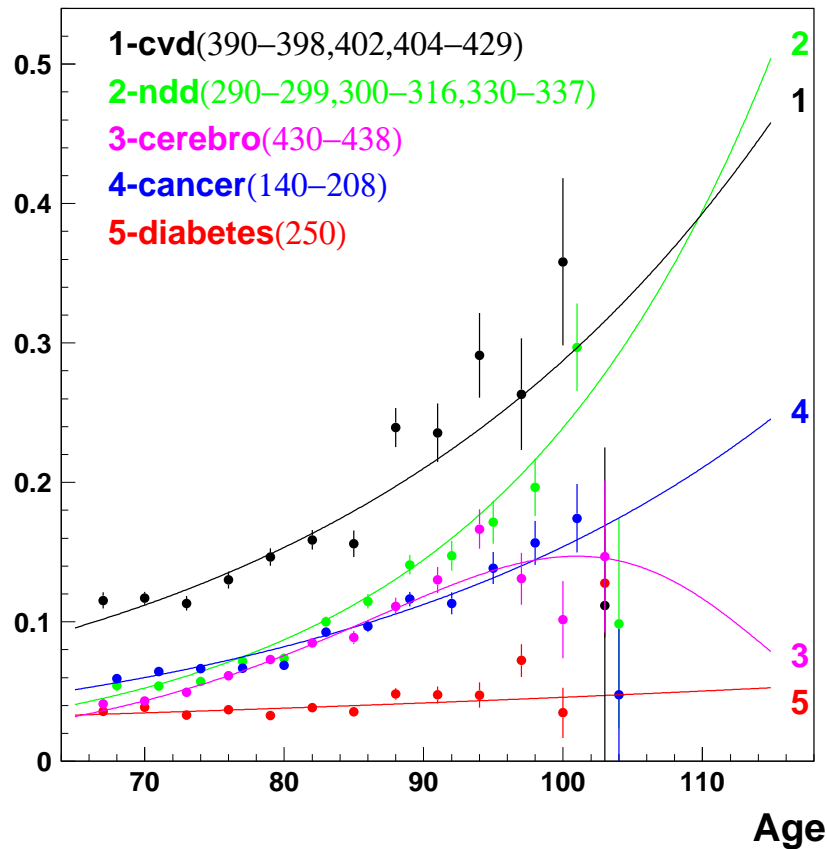
We use 1984-2001 Medicare files linked with NLTCS.

The procedure for onset identification is

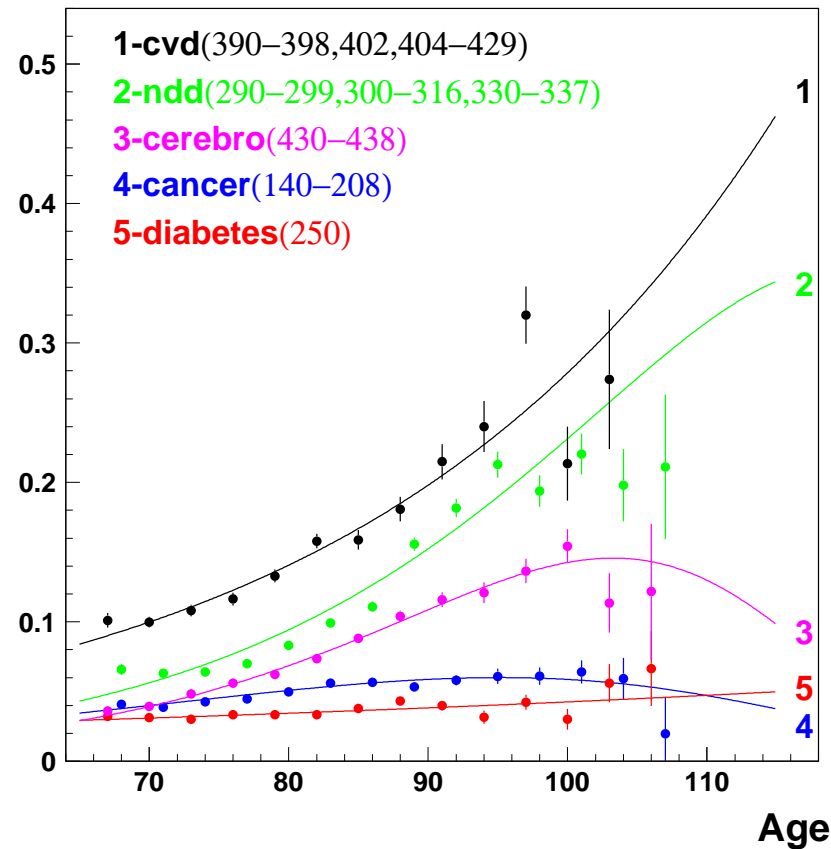
1. An individual will presumably have an onset of a certain disease in his/her observation period if there is at least one record with ICD-9-CM code corresponding to the disease on a single institutional claim or non-institutional claim on service for which beneficiary received medical care.
2. If a certain diagnosis appears in Medicare files within an initial 6-month period after enrollment into Medicare, such individual will be considered as already been chronically impaired at the time of his enrollment in Medicare.
3. Otherwise the date of the first appearance of the corresponding diagnosis will be considered as the date of onset.

# Results: Incidence rate

Incidence; Male



Incidence; Female



Age specific disease incidence: dots denote Kaplan-Meier estimations (means and SE's) of NLTCS/Medicare data for 1992-2001, lines are the generalized Gompertz model