DATA INTEGRATION AND KNOWLEDGE DISCOVERY IN BIOMEDICAL DATABASES. RELIABLE INFORMATION FROM UNRELIABLE SOURCES.

A Mitnitski[1,2*], A Mogilner[1], C MacKnight[1] and K Rockwood[1]

*[*1]Dept of Medicine, Dalhousie University, Halifax, NS B3H 2Y9*
Email: Arnold.Mitnitski@dal.ca;
[2] Dept of Computer Science, Dalhousie University, Halifax, NS B3H 2Y9
Email: Chris.MacKnight@cdha.dal.ca; Kenneth.Rockwood@dal.ca;
Email: Alex.mog@rocketmail.com;

ABSTRACT

To better understand information about human health from databases we analyzed three datasets collected for different purposes in Canada: a biomedical database of older adults, a large population survey across all adult ages, and vital statistics. Redundancy in the variables was established, and this led us to derive a generalized (macroscopic state) variable, being a fitness/frailty index that reflects both individual and group health status. Evaluation of the relationship between fitness/frailty and the mortality rate revealed that the latter could be expressed in terms of variables generally available from any cross-sectional database. In practical terms, this means that the risk of mortality might readily be assessed from standard biomedical appraisals collected for other purposes.

Keywords: CODATA, Biomedical data, Mortality, Macroscopic State variable, Ageing, Frailty, Biological constants.

1 INTRODUCTION

Scientific data must be clean and reliable. While this is the case in the majority of physical, chemical and engineering applications, biomedical data rarely possess such qualities. The very nature of biomedical objects is volatile and irregular, as are the results of biomedical assessments collected in large biomedical databases. These databases contain the results of tests which fluctuate with the patient's state, and the long term trends are difficult to distinguish from the short term fluctuations, taking into account that these databases rarely contain reliable longitudinal components. The psychological tests used for the assessments of cognitive status, which are chiefly verbal, are even more volatile depending on *who* performed an assessment, and *how* the assessment was performed (MacKnight, Graham & Rockwood, 1999). The other typical problem is the large number of incomplete records, for example, if certain tests are missing for some individuals, then deleting such records may essentially reduce the power of the ongoing calculations. Even mortality statistics, probably the most reliable type of biomedical data, are not free from error: while the date of death is usually known precisely, the date of birth can be biased. Mortality statistics are, however, believed to be of a high quality and so have been used for mathematical modeling relating rate of mortality with age since the classic works of Gompertz (1825) and Makeham (1867), and are presently used in a number of theoretical models (e.g., Strehler & Mildvan, 1960; Vaupel, Carey, Christensen, Jahnson, Yashin & Holm et al., 1998; Manton & Yashin, 1999; Gavrilov & Gavrilova, 1991; 2001) as well as in epidemiology/medicine (e.g., Greiner, Snowdon & Greiner, 1999; Jenssen, Kuo, Stokke & Hovig, 2002).

The possibility of integrating knowledge from several databases is of significant scientific and practical interest. Different databases are usually created independently, for discrete purposes, and are not linked with each other. Biomedical (epidemiological) databases generally contain information about large

numbers of individuals (health related variables: diseases, symptom and signs, physiological and psychological assessments, socio-economic variables etc.). In contrast, demographic surveys contain large amounts of data aggregated by age and sex. The databases can be linked through a common (key) variable such as age and sex. Our goal was to derive a characteristic (macroscopic state variable) from which it would be possible to suggest nontrivial predictions about the health status (e.g., adverse outcomes like probability of death). Currently, the most important predictors of mortality are chronological age (summarized in the well-known Gompertz mortality law (Manton & Yashin, 1999)) and disease states such as cancer (Edwards, Howe, Ries, Thun, Rosenberg & Yancik et al., 2002), cardiovascular disease and diabetes (Schram, Kostense, Van Dijk, Dekker, Nijpels & Bouter et al., 2002), or dementia (Wolfson, Wolfson, Asgharian, Emile, Østbye & Rockwood et al., 2001). The measures of human health accounting for morbidities and diverse environmental stresses have considerable potential within different decision-making contexts (Gold, Stevenson & Fryback, 2002; Hofstetter, & Hammitt, 2002). The indicators of disability-free life expectancy accounting for socio-professional dimensions and age are being developed (Robine, 2001; Cambois, Robine & Hayward, 2001). As discussed below, we derived a generalized variable (fitness/frailty index), which comprises all available health related information about the individual, including those usually causing more discomfort than disability. Linked with mortality data, the fitness/frailty index was proven to be a strong correlate with survival. This provides an accessible tool for appraising individual and population health status from information which is readily available in many databases.

2 METHODS

2.1 Databases

Data came from three sources: (i) The Canadian Study of Health and Aging (CSHA) (Canadian Study for Health and Aging Working Group, 1994), a representative database of elderly Canadians (n=10,267) aged 65 years old and more; (ii) The National Population Health Survey, (Swain, Catlin & Beaudet, 1999), an investigation of health, health status and the use of health services in Canada (n=81,859) across adult ages; (iii) Canadian Mortality data (Statistics Canada, 1999) – aggregated mortality data, presented in five-year intervals. The CSHA data was collected by doctors and nurses while the NPHS survey was administered in face-to-face interviews, using self reported demographic and health-related information. Variables measured health or disability and contained reasonably complete data across all age groups. These included symptoms (*e.g*. trouble with vision), disabilities (*e.g*. help in preparing meal) and disease classifications (*e.g.* high blood pressure, migraine, glaucoma). Note that these variables cross a range of severity, from items associated with an increased risk of death (*e.g.* cancer, stroke) to those that typically cause more discomfort than disability (*e.g.* dexterity, vision problems).

2.2 Deficits

Let m be the number of variables in the database. Let y_i be the i-th variable from the database ($i = 1,\ldots m$); variables can be continuous or categorical. For the purposes of our analysis we consider only binary variables or variables which can be transformed to binary. Let us introduce a threshold th_i and then transform each variable in the binary code: $x_i = 1$ if $y_i < th_i$ and $x_i = 0$ otherwise. Defined in such a way, the recoded variable will be referred to as *deficits* and the individual's record is represented as an m-dimensional binary vector. This dimension is referred to as the *embedding* dimensionality (Korn, Pagel & Faloustos, 2001). In the CSHA database we identified 92 deficits (Mitnitski, Mogilner & Rockwood, 2001) while in the NPHS database we found 38 binary deficits (Mitnitski, Mogilner, MacKnight & Rockwood, 2002b).

3 RESULTS AND DISCUSSION

3.1 "Curse of dimensionality" and relationships between the variables

The desire to assess as many characteristics as possible has an obstacle of computational tractability often referred to as the "curse of dimensionality" (Bellman, 1961). The relationships between deficits make the situation even worse, as the number of possible relationships increases much faster than the number of deficits themselves. As we mentioned, the *embedding* dimensionality is high. However, the *intrinsic* dimensionality could be much lower and even fractal (Korn, Pagel & Faloustos, 2001). In this respect the question of possible relationships between variables is of great interest: the particular patterns of the relationships between deficits are the characteristics of the disease and can be used in diagnostics. The relationships between deficits can be analyzed statistically: if the difference between the conditional probability of the occurrence of deficit X given the other deficit Y is significantly different from the unconditional probability of the occurrence of deficit X, one can say that X and Y are related (Graham, Mitnitski, Mogilner, Gauvreau & Rockwood, 1996). All the possible relationships can be represented in a relation graph (connectivity graph) (Figure 1). Vertices correspond to the deficits and the links to the statistically significant dependencies between the deficits.

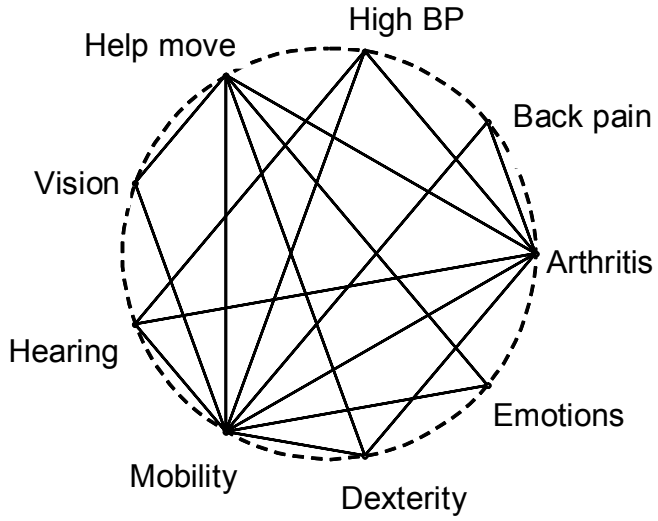

Figure 1. Connectivity (relation) graph (inter-sign synergy). Nodes (vertices) indicate the deficits (variables) and edges (links) indicate the statistically significant relationships between deficits, *i.e.* when the conditional probability of one deficit given another is statistically different ($p<0.05$;*t*-test) from the unconditional probability of the first deficit.

One can see that every deficit is related to many of the others. This indicates a very general property of the system: dependency or interrelationships is an essential characteristic which makes the system functional. For instance, in the disease groups the relationships become weaker and virtually disappeared in the severe groups (Graham et al., 1996). The abundance of the inter-variable connections prompted us to elucidate a generalized variable (macroscopic state variable) which comprises all the deficits and thus represents the whole organism rather than a particular system, organ or disease.

3.2 Fitness/frailty index (f-index) and its age trajectories

One way to construct a generalized variable (f) which accounts for all the deficits is an average of the deficits with the weight w_i ($i = 1,\ldots m$):

$$f = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} \qquad (1)$$

How to define these weights is not clear and as *a first approximation* we consider all the deficits with equal weights:

$$f = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad (2)$$

Defined in such a way, this index represents the proportion of deficits present in the individual and is called the fitness-frailty index (Mitnitski et al., 2001; Mitnitski, Graham, Mogilner & Rockwood, 2002a; Mitnitski, et al., 2002b). Two properties of the *f-index* are of particular interest. First, it was demonstrated that this variable shows the exponential kinetics with age in each of the considered databases despite the differing characteristics of the variables. Moreover, recent analyses have shown that a fitness/frailty index can be closely related to five year mortality *even when the variables that make up the index are chosen at random* (Mitnitski et al., 2001). Of course, there are certain constraints on the list of variables from which the random selection can be made, but they are chiefly technical, in the sense that the variables should be age-related (the probability of a deficit being present should increase with increasing age), and the extent to which data are missing. The *f-index* showed a simple exponential relation with age (Figure 2), indicating that the process of accumulation of deficits (damage) occurs according to an avalanche-like mechanism of random accumulation of damage (Gavrilov & Gavrilova, 1991; 2001).

$$f(t) = G + F \exp(\beta t) \qquad (3)$$

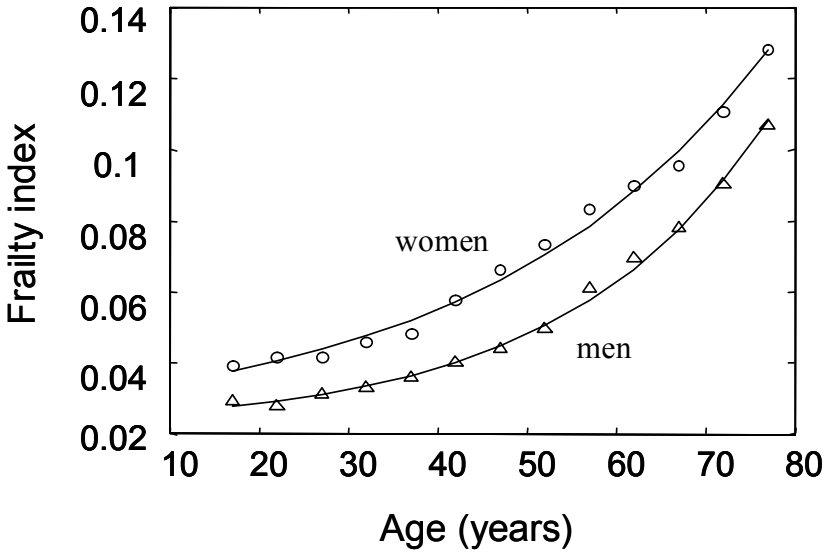

Figure 2. Accumulation of the proportion of deficits (frailty trajectories) with chronological age for men (triangles) and women (circles). Experimental data are the proportion of deficits average across the same age group. Age group is represented by the mid-point, e.g. age group from 50 to 55 years is represented by the point at 52 years. Solid lines represent fitting curves according to the two component model, Equation (3) (Mitnitski et al., 2002b).

Second, it was found that the distribution of the frailty index can be represented by a gamma density (Mitnitski et al., 2001) and for the disease groups the distribution becomes Gaussian (Figure 3).

Exponential kinetics of the frailty index and transition of its distribution from gamma to normal indicating redundancy exhaustion suggest a very general mechanism of how an organism deals with stresses. Moreover, as we will demonstrate, a simple relationship between the *f-index* and the mortality hazard exists and makes it possible to express mortality risk in terms of *f-index*.

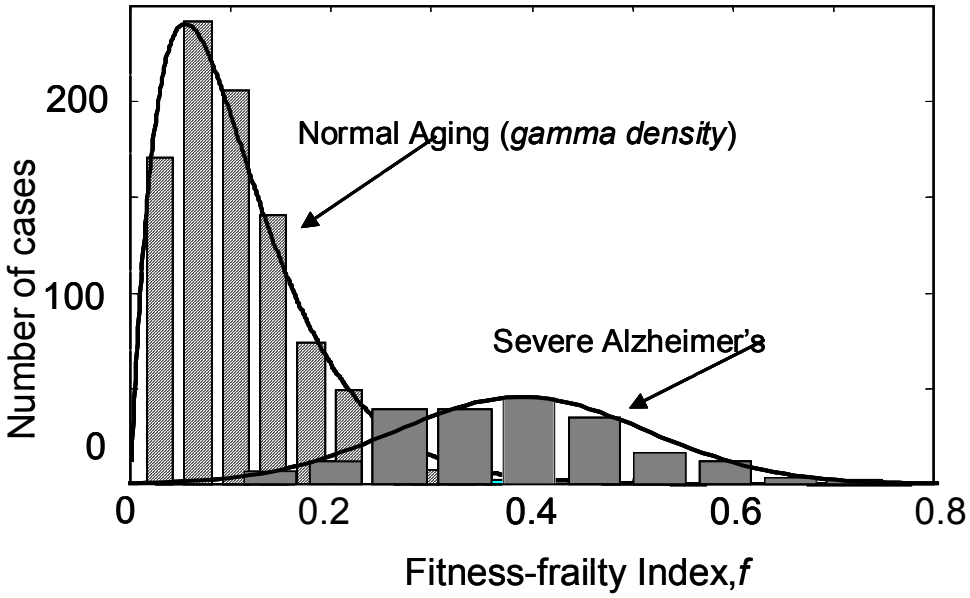

Figure 3. Transition of the statistical distribution from gamma (normal aging) to Gaussian (Alzheimer's disease).

3.3 Compensation law of fitness/frailty and biological constants

The so-called "mortality cross-over" is a well-described phenomenon that has attracted considerable attention. It refers to the observation that mortality curves tend to converge at advanced ages. This has been observed for a number of species including humans (Strehler & Mildvan, 1960; Wing, Manton, Stallard, Hames & Tryoler, 1985). Mohtashemi & Levins (2002) offered an explanation of this phenomenon based on the selection mechanism. Remarkably, but comparatively un-noticed, until perhaps more recently, it has been shown that the phenomena can be explained from the standpoint of the reliability model of mortality (Gavrilov & Gavrilova, 1991; 2001). Of great interest, from our standpoint, is that we have found analogous relationships for the accumulation of deficits. While male and female populations at the same age generally have different numbers of deficits, these differences diminish with increasing age and, in fact, crossover occurs at the age of 94 years (Figure 4) (Mitnitski et al., 2002b). Again, this result was obtained from very different data: mortality from demographic data and fitness/frailty from the biomedical datasets. This age called "species specific age in human" 95 ± 2 years (Gavrilov & Gavrilova, 2001) is an example of biological constant (Gavrilov & Gavrilova, 1991). The other biological constant is mortality (hazard) rate at this age, 0.5 1/year (Gavrilov & Gavrilova, 2001), indicating that average person at age 95 years old may evenly survive or die during the next year. The third biological constant is the fitness/frailty index corresponding to that hazard, 0.2 ± .001 (Mitnitski et al. 2002b) indicating that the damage of 20% makes the organism critically vulnerable with even chances to survive or to die. It is worth mentioning that the precise estimates of these constants will be obtained when the different population data are considered.

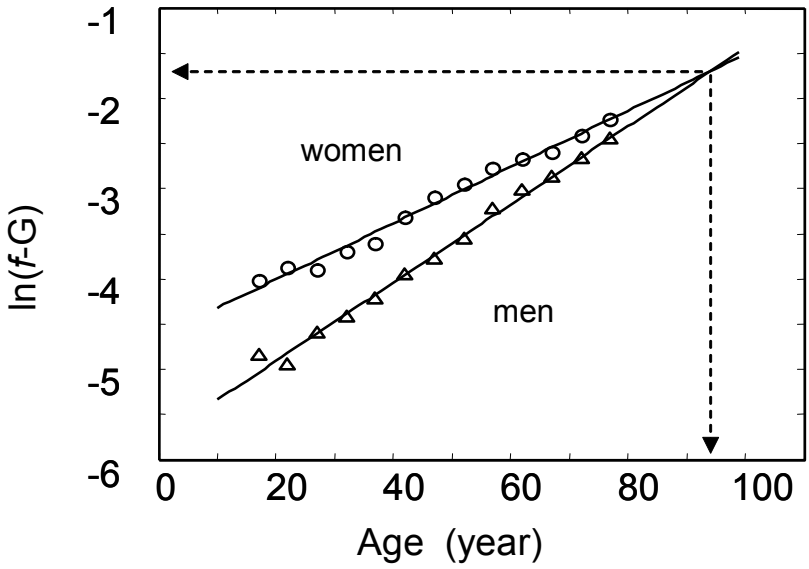

Figure 4. Compensation law of frailty (Mitnitski et al., 2002b). Logarithm of age dependent component of frailty for men (triangles) and women (circles). The least square lines have a cross point corresponding to the 0.18 of the frailty index at age of 95 years, the same age parameter for the compensation law of mortality (Gavrilov & Gavrilova, 1991; 2001).

3.4 Mortality hazard as a function of the fitness/frailty

In finding the relationship between the average mortality rate and the index, (Mitnitski et al., 2002b) we also established that the index showed a power law relationship with the mortality rate.

$$\mu = C f^{\gamma} \tag{4}$$

where C and γ are sex specific parameters (Mitnitski et al., 2002b and extended in Mitnitski, Mogilner, McKnight & Rockwood, 2002c). In Figure 5, the mortality rate is shown as a function of the fitness-frailty index in men and women, "explaining" about 99% of variance.

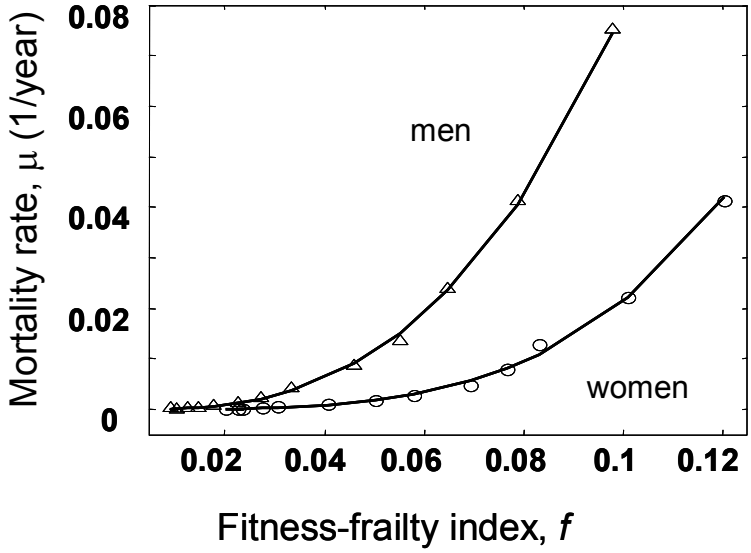

Figure 5. Power law relationships between the rate of mortality and the proportion of deficits (frailty index) for men (triangles) and women (circles). Each point represent the individuals averaged at the 5-years-old intervals, and two age groups. Arrows point to the 50-54 and 75-59 years-old men and women. Solid lines represent the least squares regressions.

This formula (Eq. 4) presents a tool for calculating the risk of mortality for a known value of the frailty index, something with great potential utility in health care and in other areas. The parameters were distinct between men and women corresponding to the fact that women on average accumulate more deficits than men at the same age, although their risk of death is lower. Thus we were not only able to confirm that the *f* index readily summarized individual differences in health status, that it did so across a wide range of ages (15-79 years) and that it worked with self-report data (NPHS), and we were also able to gain important insights into the aging rate.

3.5 Reliable information from unreliable data

As discussed above we derived a fitness/frailty index, which comprises all available health-related information about an individual in a database, including those usually causing more discomfort than disability. Being an average of many variables (deficits) obtained from various tests and examinations, each of which performed with a different degree of precision, the fitness/frailty index represents, so to say, a macroscopic state variable which indicates the general health status of the individuals. This macroscopic state variable remains relatively robust if some of its components are missing or inaccurate, as the large number of variables buffers any effect of missing or inaccurate data. A greater number of variables included in the fitness/frailty index adds to its robustness. This is a direct consequence of the large number of relationships between the deficits; roughly speaking, each deficit represents all others and this reflects not only the reliability of how an organism is functioning but also the reliability of how its status can be appraised; corresponding in a very general way with the ideas to obtaining reliable answers from a machine with unreliable components (von Neumann, 1956; von Neuman & Burks, 1966).

4 CONCLUSION

The information from different data sources was integrated from the view of the organism as a complex biomedical system with multiple connections which has a fingerprint in the occurrences of available characteristics recorded in the data sources. Complex and stochastic relationships between variables and their evolution with age can be summarized in an index variable (*f*), and that, within certain limits, it is the amount of impairment that is important, and not the type of impairment, to mean that *f* can be understood as the degree of adaptive response of the human organism to stress. Stress can come from within the system, or outside it, can be predictable or unavoidable, and can include a range of biological, social and environmental factors. Relative fitness/frailty reflects the biological and social redundancy in the adaptive response; the more the means of coping, the greater the level of fitness.

From a practical standpoint, one doesn't need to know all the variables. Any sample of them might be essential because they can all testify in an indirect way about the ability of the organism to bear stresses of different natures (environmental, social, behavioral etc.). However, taking into account as many characteristics as possible (not neglecting any of them) can be used in the assessment and monitoring of the health status of individuals and groups and contribute to the improvement of health care.

At this point we did not address the question of what is the intrinsic dimensionality of the datasets we analyzed. Nevertheless, we believe that we found an important way of representing data in one dimension by projecting the dataset on the dimension of the fitness/frailty index

In short, this relationship, if it can be replicated in other datasets, has many non-trivial consequences that need to be explored, both from a theoretical and from a practical standpoint.

4 ACKNOWLEDGEMENT

This work was funded in part by Health Canada through the National Health Research and Development Program, Grant #6603-03-1999/2640043. Chris MacKnight is supported by a New Investigator Award and Kenneth Rockwood by an Investigator Award, each from the Canadian Institutes of Health Research. Kenneth Rockwood is also supported by the Dalhousie Medical Research Foundation as the Kathryn Allen Weldon Professor of Alzheimer Research.

5 REFERENCES

Bellman, R. (1961) *Adaptive Control Processes: A Guided Tour*, New Jersey: Princeton University Press.

Cambois, E., Robine, J.M., & Hayward, M.D. (2001) Social inequalities in disability-free life expectancy in the French male population, 1980-1991. *Demography* 38, 513-524.

Canadian Study of Health and Aging Working Group (1994) The Canadian Study of Health and Aging: study methods and prevalence of dementia. *CMAJ* 150, 899-913.

Edwards, B.K., Howe, H.L., Ries, L.A., Thun, M.J., Rosenberg, H.M., Yancik, R., Wingo, P.A., Jemal, A., & Feigal, E.G. (2002) Annual report to the nation on the status of cancer, 1973-1999, featuring implications of age and aging on U.S. cancer burden. *Cancer* 94 (10), 2766-2792.

Graham, J.E., Mitnitski, A.B., Mogilner, A,I, Gauvreau , D., & Rockwood, K. (1996) Symptoms and signs in dementia: synergy and antagonism. *Dementia* 7, 331-335.

Gavrilov, L.A., & Gavrilova, N.S. (1991) The Biology of Life Span. A Quantitative Approach, V. Skulachev, V. (Ed.), NewYork: Harwood Academic Publishers.

Gavrilov, L.A., & Gavrilova, N.S. (2001) The reliability theory of aging and longevity. *J.Theor Biol* 213, 527-545.

Gold, M.R., Stevenson, D., & Fryback, D.G. (2002) HALYS and QALYS and DALYS, on my: similarities and differences in summary measures of population Health. *Annu Rev Public Health* 23, 115-134.

Gompertz, B. (1925) On the nature of the function expressive of the law of human mortality and a new mode of determining life contingencies. *Phil Trans R Soc* 2, 513-585.

Greiner, P.A., Snowdon, D.A., & Greiner, L.H. (1999) Self-rated function, self-rated health, and postmortem evidence of brain infarcts: findings from the Nun Study. *J Gerontol B Psychol Sci Soc Sci*,54, S219-22.

Hofstetter, P., & Hammitt, J.K. (2002) Selecting human health metrics for environmental decision-support tools. *Risk Anal* 22, 965-983.

Jenssen, T.K., Kuo, W.P., Stokke, T., & Hovig, E. (2002) Associations between gene expressions in breast cancer and patient survival. *Hum Genet* 111, 411-20.

Korn, F., Pagel, B-U., & Faloustos, C. (2001) On the "Dimensionality Curse" and Self-Similarity Blessing", *IEEE Trans, Knowledge and Data Eng. 13* (1), 96-111.

MacKnight, C., Graham, J., & Rockwood, K. (1999) Factors associated with inconsistent diagnosis of dementia between physicians and neuropsychologists. *J Am Geriatr Soc.* 47, 1294-1299.

Makeham, W.H. (1867) On the law of mortality. *J Inst Actuaries*. 13, 325-358.

Manton, K.G., & Yashin, I.A. (1999) Inequalities of life: statistical analysis and modelling perspectives. Chapter 4. In Sauvarin-Dugerdil, C. (Ed.) *Age between Nature and Culture,* Geneva Switzerland: University of Geneva.

Mitnitski, A.B., Mogilner, A.J., & Rockwood, K. (2001) Accumulation of deficits as a proxy measure of aging. *The Scientific World Journal* 1, 323-36. Retrieved November 10, 2001 from: http://www.thescientificworld.com

Mitnitski, A.B., Graham, J.E., Mogilner, A.J., & Rockwood, K. (2002a) Frailty, fitness and late-life mortality in relation to chronological and biological age. *BMC Geriatrics 2* (1) Retrieved December 22, 2002 from the BioMed Central website: http://www.biomedcentral.com/1471-2318/2/1

Mitnitski, A.B., Mogilner, A.J., MacKnight, C., & Rockwood, K. (2002b) The accumulation of deficits with age and possible invariants of aging. *The Scientific World Journal* 2, 1816-1822. Retrieved July 10, 2002 from: http://www.thescientificworld.com

Mitnitski, A.B., Mogilner. A.J., MacKnight. C., & Rockwood, K. (2002c) The mortality rate as a function of accumulated deficits in a frailty index. *Mech Ageing Dev* 123, 1459-1462.

Mohtashemi, M., & Levins, R. (2002) Qualitative analysis of the all-cause black-white mortality crossover. *Bull Math Biol* 64, 147-173.

von Neumann, J. (1956) Probabilistic Logic and the Synthesis of Reliable Organisms from Unreliable Components. In Shannon, C.E. & McCarthy, J. (Eds), *Automata Studies*. Princeton, NJ: Princeton University Press.

von Neumann, J. & Burks, A.W. (1966) *Theory of Self-reproducing Automata*. Urbana, IL: University of Illinous Press.

Robine. J.M. (2001) A new biodemographic model to explain the trajectory of mortality. *Exp Gerontol* 36, 899-914.

Rockwood, K., Hogan, D.B., & MacKnight, C. (2000) Conceptualization and measurement of frailty. *Drugs Aging* 17, 295-302.

Rockwood, K., Stadnyk, K., MacKnight, C., McDowell, I., Hébert, R., & Hogan, D.B. (1999) A brief clinical instrument to classify frailty in elderly people. *Lancet* 353, 205-206.

Schram, M.T., Kostense, P.J., Van Dijk, R.A., Dekker, J.M., Nijpels, G., Bouter, L.M., Heine, R.J., & Stehouwer, C.D. (2002) Diabetes, pulse pressure and cardiovascular mortality: the Hoorn Study. *J Hypertens* 20(9), 1743-1751.

Statistics Canada (1999) Health Statistics Division. Mortality- Summary List of Causes 1997, Catalogue no 84F0209XIB. Retrieved January 10, 2001 from the Statistics Canada website: http://www.statcan.ca/english/IPS/Data/84F0209XIB.htm

Strehler, B.L., & Mildvan, A.S. (1960) A General Theory of Mortality and Aging. *Science* 132, 14-21.

Swain, L., Catlin, G., & Beaudet, M.P. (1999) The National Population Health Survey: its longitudinal nature. *Health Reports* 10, 69-82.

Vaupel. J.W., Carey, J.R., Christensen, K., Jahnson, T.,E., Yashin, A.,I., Holm, N.V., Iashine, I.A., Kannisto, V., Khazaeli, A.A., Liedo, P., Longo, V.,D., Zeng, Y., Manton, K.G., & Curtsinger, J.W. (1998) Biodemographic trajectories of longevity. *Science* 280, 855-860.

Wing, S., Manton, K.G, Stallard, E., Hames, C.G., & Tryoler, H.A. (1985) The black/white mortality crossover: investigation in a community-based study. *J Gerontol* 40, 78-84.

Wolfson, C., Wolfson, D., Asgharian, M., Emile, C., Østbye, V., Rockwood, K., & Hogan, D. (2001) A re-evaluation of the duration of survival from the onset of dementia. *New Engl J Med* 344, 1111-1116.