DATA INTEGRATION AND KNOWLEDGE DISCOVERY IN BIOMEDICAL DATABASES. A CASE STUDY.

A Mitnitski[1,2*], A Mogilner[3], C MacKnight[2] and K Rockwood[2]

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

INTRODUCTION

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 et al., 2002), cardiovascular disease and diabetes (Schram et al., 2002), or dementia (Wolfson et al., 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.

METHODS

Databases
Data came from three sources: (i) The 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 NPHS (National Population Health Survey, Swain et al., 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).

Deficits
Let *m* be the number of variables in the database. Let y_i be the *i*-th variable from the database ($i = 1,…m$); 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 et al., 2001). In the CSHA database we identified 92 deficits (Mitnitski et al., 2001) while in the NPHS database we found 38 binary deficits (Mitnitski et al., 2002).

RESULTS AND DISCUSSION

"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 et al., 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 et al., 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.
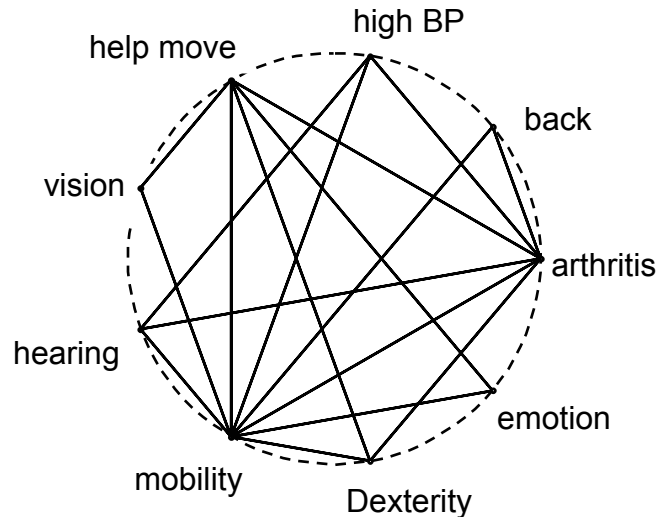

Figure 1. Connectivity graph (inter-sign synergy [Mitnitski et al. 2002a]). Nodes indicate the deficits (codes correspond to the deficits in Mitnitski et al., 2002a) and edges indicate the statistically significant relationships between deficits, *i.e.* when the conditional probability of one deficit 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.

Fitness/frailty index (ff-index) and its age trajectories
One way to construct a generalized variable (*ff*) which accounts for all the deficits is an average of the deficits with the weight w_i ($i = 1,…m$):

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

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

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

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, 2002a,b). Two properties of the *ff-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 *ff-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 damaged (Gavrilov & Gavrilova, 1991, 2001).

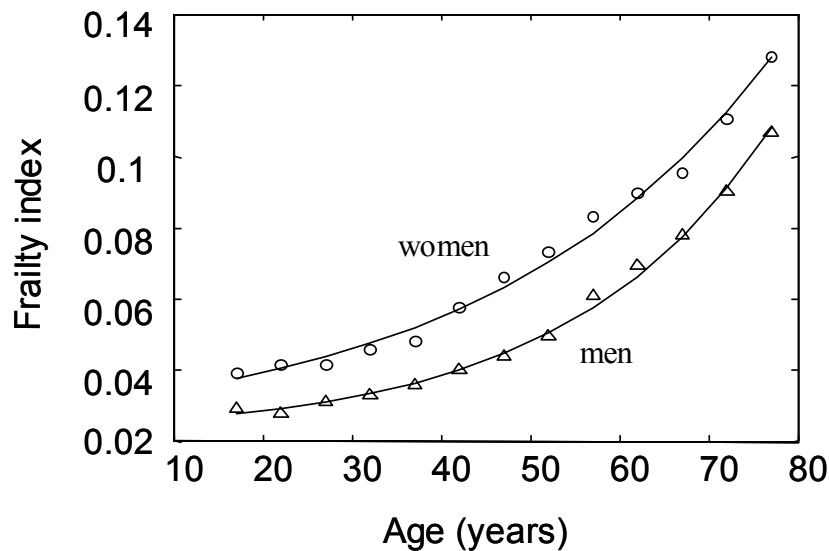$$ff(t) = G + F \, exp(\beta \, t) \qquad (3)$$

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 *ff-index* and the mortality hazard exists and makes it possible to express mortality risk in terms of *ff-index*.
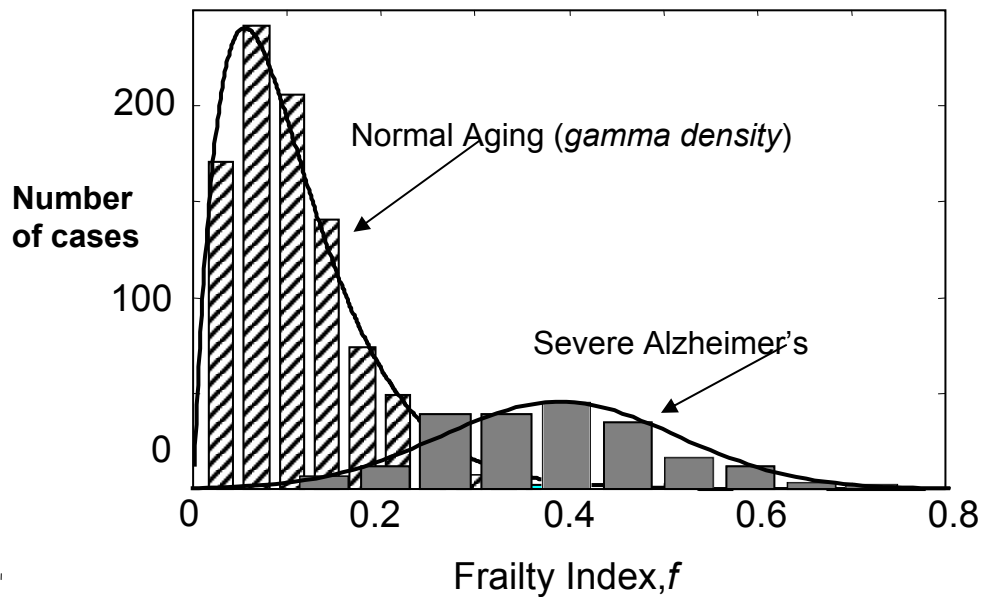

Figure 3. Transition of the statistical distribution from gamma (normal aging) to Gaussian (Alzheimer's disease).

Compensation law of fitness/frailty
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 (Strehler & Mildvan, 1960) including humans. 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 95 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.
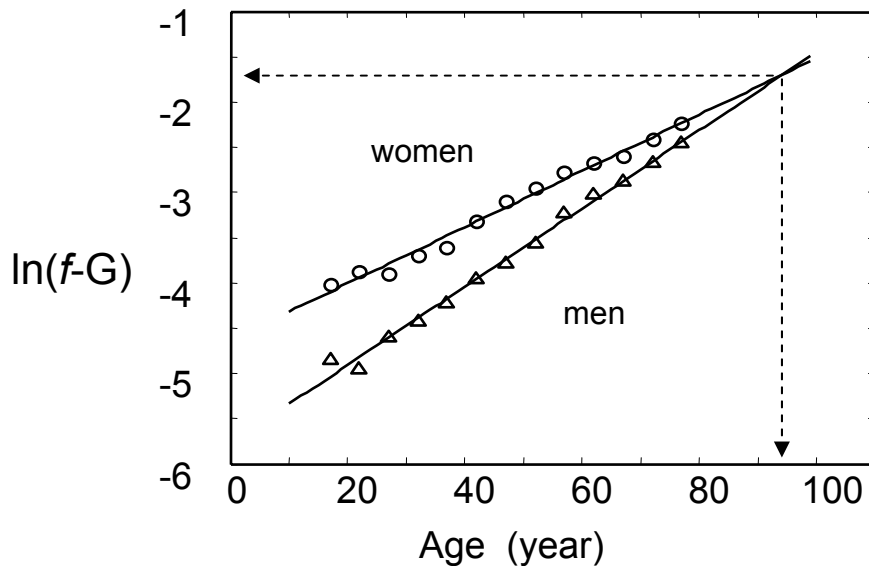

Figure 4. Compensation law of frailty [Mitnitski et al., 2002b]. Logarithm of age dependent component of frailty for men (triangles) and women (circles). The least square lines have a cross point corresponding to the 0.18 of the frailty index at age of 95 years, the same age parameter for the compensation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

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.

$$\mu = C\,ff^{\gamma} \tag{4}$$

where C and γ are sex specific parameters (Mitnitski et al., 2002b and extended in Mitnitski et al., in press). 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. This formula 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 *ff* 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.
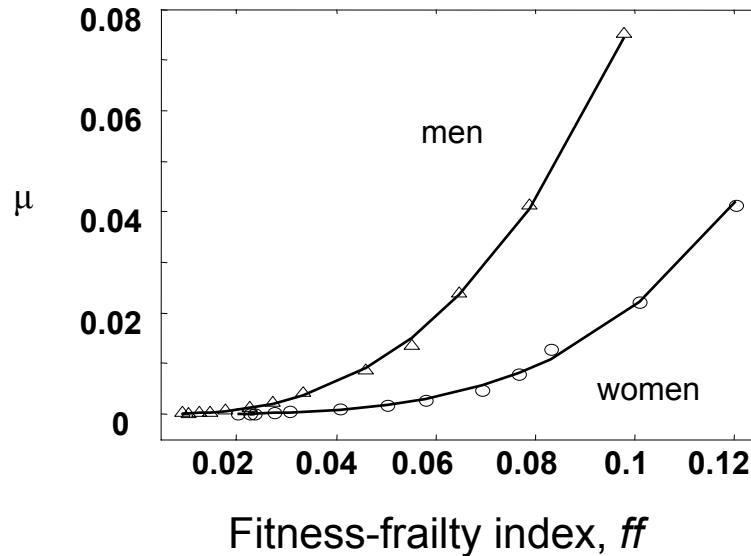
Figure 5.   Power law relationships between the rate of mortality and the proportion of deficits (frailty index) for men (triangles) and women (circles). Each point represent the individuals averaged at the 5-years-old intervals, and two age groups. Arrows point to the 50-54 and 75-59 years-old men and women. Solid lines represent the least squares regressions.

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

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 (ff), and that, within certain limits, it is the amount of impairment that is important, and not the type of impairment, to mean that ff 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 the stresses of the different nature (environmental, social, behavioral etc.). However, taking into account as many characteristics as possible (not neglecting

any of them) can be used in the assessment of the health status in individuals and groups and in monitoring 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 dataset on the dimension of the fitness/frailty index

REFERENCES

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

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 BK, Howe HL, Ries LA, Thun MJ, Rosenberg HM, Yancik R, et al., Annual report to the nation on the status of cancer, 1973-1999, featuring implications of age and aging on U.S. cancer burden. Cancer 2002;94(10):2766-92

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., and Gavrilova, N.S., 1991. The Biology of Life Span. A Quantitative Approach, V. Skulachev, V. (Ed.), Harwood Academic Publishers.

Gavrilov LA, Gavrilova NS. The reliability theory of aging and longevity. J.Theor Biol 2001; 213, 527-545

Korn F, Pagel B-U, Faloustos Ch. On the "Dimensionality Curse" and Self-Similarity Blessing", IEEE Trans, Knowledge and Data Eng. 2001;13,1: 96-111.

Manton KG, Yashin IA. Inequalities of life: statistical analysis and modelling perspectives. Chapter 4 in: Age between Nature and Culture (Sauvarin-Dugerdil C., Ed.) University of Geneva, Geneva Switzerland. 1999.

Mitnitski AB, Graham JE, Mogilner AJ, Rockwood K Frailty, fitness and late-life mortality in relation to chronological and biological age. BMC Geriatrics, 2002; 2:1.

Mitnitski A, Mogilner A, Rockwood K. Accumulation of deficits as a proxy measure of aging. The Scientific World 2001; 1: 323-36.

Mitnitski AB, Mogilner AJ, MacKnight C, Rockwood K. The accumulation of deficits with age and possible invariants of aging. The Scientific World 2002a;2: 1816-1822.

Mitnitski AB, Mogilner AJ, MacKnight C, Rockwood K. The mortality rate as a function of accumulated deficits in a frailty index. Mech Ageing Dev 2002b (in press)

Rockwood, K., Hogan, D.B., and 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 MT, Kostense PJ, Van Dijk RA, Dekker JM, Nijpels G, Bouter LM, Heine RJ, Stehouwer CD. Diabetes, pulse pressure and cardiovascular mortality: the Hoorn Study. J Hypertens 2002; 20(9):1743-1751.

Statistics Canada., 1999. Health Statistics Division. Mortality- Summary List of Causes 1997, Catalogue no 84F0209XIB.

Swain L, Catlin G, Beaudet MP1999. The National Population Health Survey: its longitudinal nature. Health Reports 10, 69-82.

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