Informal Discussion Transcript
General Session V: The Nitty-Gritty of the Human Mortality Database

Presented at the Living to 100 Symposium
Orlando, Fla.
January 4–6, 2017

DALE HALL: Good morning, everyone. I'm Dale Hall, managing director of research at the Society of Actuaries. Thanks for joining us again this morning. One of the things I really enjoy about conferences like this is the networking we get to do over hallway conversations and breakfast, and then the problem is trying to get everyone into the General Session here for presentation.

The session that we have this morning is entitled "The Nitty-Gritty of the Human Mortality Database." We're privileged to be joined by Magali Barbieri, who is the associate director of HMD, stationed at UC Berkeley, and also part of the French National Institute for Demographic Studies.

It's been a real privilege of the SOA to work more and more with HMD over the past year on different projects that we've been sponsoring, so that actuaries all around the world can use information from the valuable population databases that they create. I had a chance to meet with Magali and her staff at Berkeley about a month ago—a pretty impressive group of people that are there to keep the momentum going with HMD analysis in the future.

This is intended to be a presentation by Magali, but we realize there's a lot of charts and graphs and methods and things that will be described, so please feel free to raise a hand or ask a clarifying question, even during the presentation. I don't necessarily want to have to wait until 10 minutes before the end of the session, where we have to go back to something that was on slide 2 or 3. So we do have microphones here, or if you can flag, raise a hand, or maybe run to the microphone or ask a question, and we can get it repeated and clarified, that would be helpful for the whole session.

A couple of housekeeping items: We do have Enrolled Actuary credit forms available for this session, so if you need to sign off on those, find one of us, and we can help point you in the right direction. And then we do have an evaluation form. We'd like you to fill that out; give us some feedback on the session as it completes.

So with that, I'm going to turn it over to Magali. Please help me welcome Magali Barbieri.

MAGALI BARBIERI: Thank you, Dale, and thank you to the SOA in general for, first of all, very generous support to the project—financial support last year and then again this year, and

hopefully this is the beginning of a collaboration that's going to last for the next few years. And thank you also for bringing me here, for inviting me to participate to this meeting.

As you will see very shortly, the HMD involved a very time-consuming process to produce all of this data that we publish in our website. It involves the work of a bunch of people, and it's always a pleasure to see that our work is appreciated and used extensively, and I must say that your community represents a really large share of all the HMD users. As I mentioned a couple of days ago, we now have about 40,000 registered users who access our database on a regular basis, and about a quarter of them are from your profession. So thank you for the support that gives sense to the work we do, gives some meaning to this hard work that we've been doing over the past 15 years. And, again, it's a great opportunity for me to have this platform to exchange with you and better understand your needs, so as to further improve the HMD.

With that, I'm going to start this presentation, which again concentrates on issues of challenges and methods that have been developed over the years by a large number of people within the HMD and outside of the HMD as we entered into many conversations about modern demographic techniques with our community in general.

So let me start here. So, again, I will emphasize what Dale just mentioned: Don't hesitate to interrupt me by coming here to the microphone if you have questions about a particular slide or a particular concept that I'm explaining. We'll keep the more fundamental questions for the end, for the general discussion at the end, but if there is any clarifying point that you need me to get into, don't hesitate to interrupt.

So I made a general presentation a couple of days ago about the HMD. For people who are not so familiar with the database, I'm going to quickly review here how much of this I need to go over again this morning before I really start discussing the methods, so please raise your hands if you have never heard about the HMD before. Okay, I can't even see if there are any hands raised. Yeah, there were a couple.

DALE HALL: About 10 or 15.

MAGALI BARBIERI: And among the people who know about the HMD, how many have actually downloaded data, have gotten to the point not only of looking—? Okay, good. So I'm

going to go very quickly about the generalities, and I will try to spend most of my time going into the details of the process—the whole production process of the HMD.

So very quickly, basically the HMD is a database of life tables. These life tables are provided by single calendar years, at quite a detailed level of single years of age up to the age of 130, for 38 countries for extended periods of time. And the HMD provides not only these life tables, which we have reconstructed, but also the original data—the original demographic statistics from which these life tables have been constructed—as well as an extensive documentation to ensure the maximum transparency in terms of the process going from the original data to the life tables.

The guiding principles of the HMD from the very start: So the HMD was launched in 2002 after several years of work by its original funders. And from the start, the guiding principles were comparability, so we thrive to produce the data in a very standardized manner over time and across countries, so all of the life tables are perfectly comparable to the extent that the original data allow it. Another guiding principle is accessibility; the data have been and will continue to be available for free, thanks to the support of many organizations. We also try to make the data as flexible as possible; we provide the output data in a number of formats. Reproducibility, which I mentioned before, is an important one. And then finally, quality control, and I think that's one of the reasons why the HMD is so popular. It's because it is recognized that our major efforts to control the reliability and the accuracy of the data both at the very beginning, the input data, but also the data we produce ourselves, and I will mention a number of things we do to make sure the data is of the highest quality.

So for additional general information about the HMD, I refer you to this article that we published last year, which gives a very detailed overview of the database. And don't hesitate to contact me to ask for a copy of this paper that I'll be happy to send.

So the HMD provides data for, I mentioned it, 38 countries as of today. We've investigated the possibility to add more countries. We've looked at other countries that we decided not to include because of quality issues. These 38 countries were selected because currently their data are of the highest quality. They pass the bar of a number of standards we require, and they also are

available—the original data are available in—they are detailed enough that they can be processed through the HMD machine.

So as you notice, most of, if not all of, these countries—except for a couple, maybe—are all high-income countries: from Europe, from all parts of Europe, although not all European countries are included. So, for instance, we investigated the inclusion of Romania, Croatia and Moldova and decided against including these countries at this point in the HMD, but most of Europe, including Eastern European countries; North America; Australia; and New Zealand; and then Chile, which is the only country from Latin America that we've included so far, but we have misgivings about that, and those of you who have been looking at Chile will see that the series stops in 2005 because of serious data concerns about a recent census in this country; as well as three countries in Asia: Japan, of course, but also Taiwan and Israel, if we can consider Israel to be an Asian country.

The series are extended as far back as the data make it possible, so again, in terms of availability and quality. So for Sweden, which is the country with the longest time period, we have about 250 years' worth of data, and for Chile, which has the shortest, we have less than 15 years.

We only produce cohort life tables, so the life tables that are available in the HMD are period life tables, though we have cohort life tables for a number of countries—basically, countries which have at least one extended cohort. So what it means in practical terms [is] that we have at least 100 years or 110 years' worth of data. So we will continue constructing these cohort series as time goes on, and we're updating the HMD on a regular basis. So as we're adding years of data, we'll be able to increase the number of countries for which we have cohort information.

So ideally, the data that we need to construct HMD series are very detailed. So the mortality data we need are date counts, not only by calendar year of occurrence, by sex and single year of age, but also by year of birth by cohort. The birth dates that we use are live births by sex, and because of new methods that we've introduced very recently, we also use live births by month. And the population data we need, ideally again, are January 1 estimates of population by sex and single year of age up to the highest age possible.

So this perfect data is only available for a relatively short time period. From these ideal

data, the construction of life tables is pretty basic. We first construct death counts and exposure counts by Lexis triangle—and then we'll get back to a short description of what we mean by Lexis triangle, but basically death counts by year of occurrence, single year of age and cohorts, and exposure counts for the corresponding population. And the methodological steps we take are first to construct these exposure and death counts by Lexis triangles, then compute the death rates from the ratio of the death to the exposures, and then from the death rates, we compute complete life tables, and from the complete life tables, the abridged life tables—five-year age group life tables.

So this is a Lexis diagram; I'm going to assume that you all know what a Lexis diagram is. So time is on the *x*-axis, age on the *y*-axis, and then what makes it a Lexis diagram is the cohort information that's added and that runs on diagonals throughout this graphic. So, for instance, here you have a representation of the itinerary of the 2011 cohort as it ages through time. So this is what we mean by a Lexis triangle. So a Lexis triangle is basically, on this figure, a shape that represents a unique combination of calendar year, age, and cohort. So a cohort is a group of people who are born during the same calendar year.

So we compute the death rates in each of these triangles, and then we arrange the triangles to compute death rates by single year of age, either along the period or along the cohort, to build our life table. So that's the basic principle.

So when we have population or death counts by single years of age, if we don't have the detail about the cohort, we need to desegregate these numbers into an upper triangle. Is there a pointer? Is that it? Okay. So this is what we call the upper triangle, and this is what we call the lower triangle. And the first step that we take in the HMD is to compute, if they're not available directly, we compute death counts and exposure counts within each of these triangles.

But there are very few countries and time periods for which we have this perfect data, so mostly these are the countries—mostly in Scandinavia, in Northern Europe—for which we have this most detailed data available. In the vast majority of our countries and for much of the time period, even for these Scandinavian countries, we don't have that level of detail, and that's where we've spent a lot of time devising methods or selecting methods. Usually, we don't develop our own methods. We test everything that is available, that has been published, that has been

substantiated, and then we make decisions, by a number of validation processes, about what's the best method that we can use. So we've developed a whole set of methods to deal with data which are not available at this level of detail.

And we faced a number of data challenges. So as I told you, we select countries which have the highest-quality data currently, but we try to extend the series as far back as possible with the data that's available, and of course, when you go back to the 19th century or even the 18th century for some countries, the quality of the data cannot be expected to be of the same level as they are now. So—and they're also not at the same level of detail—so we devised a number of methods to check the quality of the original data and, if we're satisfied with the quality, to massage the data in a way that can be easily processed through the HMD machine.

So the types of data challenges we face are numerous, and I've tried to summarize them here. So there are issues of availability. As the Canadians in the room know, there are issues of publication delays for Canada, which we've been discussing extensively recently with our Canadian colleagues. The Canadian data series only goes up to 2011 currently. Statistics Canada has released the 2012 data but not yet the 2013 data, and even for the 2012 data, there are issues of access. There are problems of details in the data that's available, so for mortality data, for instance, it's actually rare for a historical period to have data cross-classified both by single year of age and by birth cohort. So at best, we have single year of age data, but often—especially, again, when we go back in time—the data is only available in five-year age groups and sometime even in 10-year age groups, so we have to disaggregate the death counts or the population counts.

There are issues of definitions in terms—for instance, of live birth or reference population. We also have issues when country borders have been modified, and we have a process to deal with that, because this creates discrepancies between the numerator and the denominator in terms of the population that's covered.

There are, of course, last but not least, issues of reliability in terms of registration coverage or the proportion of deaths or population at unknown ages. And then, even for contemporary populations, we have issues of age misstatement, and in many cases, including the U.S., we have concerns about age, especially age overstatement at older ages, and I don't think in this meeting

there has been a discussion about that, but there have been other meetings of the SOA where this has been discussed.

So, again, all of the methods we've developed have been designed to target these different issues, but in a way that's highly standardized. So if we have the same problem for different time periods, for different countries, we apply systematically the same method, because again, one of the principles of the HMD is comparability. So each of these methods might not be the best method for that country, for that time period, but it's the best method overall for all of the countries' time periods in that particular situation.

So the HMD is a multistep process. The first step is, of course, to gather the raw data, and nowadays, most of the data is available in electronic format. But as we go back in time, we've had to digitize published statistical volumes, and if we go back even further in time, we've had to work in the archives of many of the statistical offices of the countries we've been working on. So it's been a very time-consuming process just to collect the data that we've needed, and then also to check carefully all of these data.

Then the next step is to format the data, because again, we run a very centralized process to go from the input data to the life tables. We have a single set of computer programs that are maintained at UC Berkeley—have been developed, of course, with our collaborators at the Max Planck—so there is a single set of programs that's applied with a number of routines which are used for particular situations, depending on the time period and the country. But there is a single set of programs that are run and which require the data to be formatted in a particular way.

And then there is a first set of programs that estimates death counts and exposure counts by Lexis triangles, and there is another set of verifications at this stage. That's also the stage at which we carry out some territory adjustments when there have been changes in country borders.

The next step is to calculate death rates by Lexis triangles. This is very straightforward: This is the ratio of the deaths to the population or to the exposures within each Lexis triangle. And then from the death rates—which we actually smooth at higher ages, and I'll explain more about that in a few minutes—we construct the complete life tables, and from the complete life tables, the abridged life tables.

And then a new set of verifications, both for internal consistency and also we carry out some external validation to the extent that we have life tables that are available either from national statistics offices or from academics. And here I want to emphasize the fact that we've developed over the years a very strong relationship with statisticians in statistical office, and much of the work we do and the quality that we strive to achieve would not be possible without their help, because of their degree of knowledge of the data and also their support in providing data in more detail than what's publicly available. And I must say that we've been able to rely on very competent people in most of these countries.

Then after this step of external validation, we complete the documentation, which we update each time we update the data series, and everything. So we have a very detailed checklist for the country specialists. So the 38 countries are divided among, we have a team of about 10 people, and the countries are divided among all of these 10 people, who have developed a particular experience with their countries over the years and particular relationships with the data producers in the national statistics offices—because all of the data we used, I didn't mention that, but all of the data we used are official demographic statistics. And then the country specialists have to fill out these very extensive checklists, and all of the work is reviewed by Vladimir Shkolnikov at the Max Planck and by myself at Berkeley for all of the countries. So all of the countries are verified by the two of us, and then when we're satisfied—and often there is a back-and-forth process with the country specialist and even with the national statistics office when they are idiosyncrasies or anomalies in the data—and then finally when everyone is happy about the results, we publish the data, and sometimes with warnings for our users.

So this is an example of the methods protocol, which is described in much detail, and I won't have time today to go through all of these details, but I'm just going to give you an overview, an idea of the kind of approach we're following, but the methods protocol is about 80 pages long. This is the fifth version, because we try to improve the methods as time goes on, and research on these different demographic techniques is published, and we're currently in the process of transitioning from this version 5 to a version 6, and the version 6 will include two new methods, which I will briefly mention during my talk.

So we start with the birth data, and the way we use the birth date is to verify, so throughout our internal consistency checks, we verify the deaths and populations for the first year in comparison with the birth data, and we also use the birth data—and you'll see that shortly—to estimate the size of individual cohorts from birth until the first time when we have data for that birth cohort. And then finally—and this is one of the two new methods we're implementing in version 6—we use the birth data and more specifically the birth-by-month data to adjust for calendar years when there is a non-uniform distribution of birth. And we'll see that's very important, and we have not been doing that until now, which creates some distortion in the death rate series, and I'll give you a graphical example that's very straightforward.

Next, we reformat the death counts. Deaths come in a variety of formats and with various issues, including deaths of unknown age that we redistrict proportionately. And here our goal is to go from the input death counts to deaths by deaths Lexis triangles, and we also have to deal with open aging intervals, because only for very few countries and for the most recent time periods [do] we have deaths up to the highest possible age, and even when that's the case—for instance, for the U.S.—there are some highly unreliable results. We've had in the U.S., in the past, people who died at age 137, for instance, which is not very reasonable, and that's also the case for a number of other countries. So we typically group the deaths that we get into an open age interval, or sometimes they come in this format especially for longer time periods, and we redistribute the deaths from the open age interval into Lexis triangles also.

So this is the type of Lexis shape in which the data has been provided over the years. So we have not only differences between countries, but within countries depending on the time period, the data come in various shapes: by cohort, by period, by single year of age, by five-year, 10-year age groups and so on.

So the very first step is to split the deaths from age groups to single years of age, and to do that, we just use a spline, which is an equation of this shape. Okay, so I'm going to go quickly over that. So I'm going to show you a number of equations, but I'm not going to go into details about that. I'm just going to give you the intuitive idea behind each of these methods. So we use a spline to redistribute deaths from age groups to single years of age, and then we use a regression that's

been developed carefully from data for which we have highly detailed information to redistribute the deaths from single years of age into Lexis triangles.

So this is basically the data that's been used to develop the regression. So for these countries, we have deaths cross-tabulated for extensive time periods, cross-tabulated by single years of age and birth cohorts up to the highest age. So we use the information for these countries to develop a model that's based on a number of indicators. So the equation has—this is an example for males; we have a slightly different equation for females—so we use information on the share of births in the lower triangle, as the respective size of the birth cohort has an impact on how to allocate the death counts by single years of age into the two Lexis triangles. We also include information about the influenza epidemic of 1918 and 1919, which was very seasonal, so it had a big impact on where the deaths should be allocated. We use information on the level of infant mortality, which is an indicator of the overall level of mortality, as this also has an impact on the allocation of deaths in the lower or upper triangle, and we have an interaction, so we use specific coefficients for the first year of life and for the next age interval, age one. And then, finally, we have a particular adjustment for countries with very low levels of mortality.

So this is just to show you how complex some of these estimations are. Again, for women, we have a very similar equation with the same infant mortality indicator and the same birth distribution indicator. I'm not going to show you that; we can discuss that later, but these are validation graphs for this method.

Then the next step for the death counts is to redistribute the deaths from the open age interval into Lexis triangles, and here we use the method that's been developed by Väinö Kannisto. So the intuitive explanation is that we use the information we have for other age groups to determine what the mortality curve should look like at these higher ages for which all of the data are aggregated into a single age group.

So at this point, we have deaths by Lexis triangles up to age 130, and we reconcile numerators and denominators when there have been territorial changes at that point, using coefficients which vary by age. We also use this method of territorial adjustment in situations where we do have countries with no territorial border change but changes in population definitions

in a way that we cannot apply our typical set of methods to construct intercensal estimates.

So this is an example of Poland between 1960 and 2014. Poland joined the European Community in 2004, and the European Community provides funding depending on population size to each country, so each country has a strong incentive to have a large population. And so Poland, which had revised its way of counting its reference population based on permanent residency in previous censuses, decided to switch to a new definition of population based on usual residence. So it's not even de facto, because in these usual residents are counted people who have left the country but supposedly for short time periods, and again, as soon as Poland joined the EU in 2004, there were huge migration waves outside of the country, of Polish people moving to other European countries to work there. And so we had this issue of official population estimates here that were completely inconsistent with previous estimates or future estimates. So this is 2002 and 2011, and so we could not use our typical method for intercensal estimate because of the large migration swings, and so we used this territorial adjustment to adjust here. And so the estimates we actually use are these green values.

So next, we have to do the same kind of work to produce exposures by Lexis triangles, and so we work with the population data, and we apply a number of methods, and the methods we are applying depend not only on the details in the data that's available, but also on the age. So except for very few countries which we find are completely reliable up to the highest age and which provide data in much detail—which are Sweden and a few other Northern European countries—basically for all countries, we do not use the population at ages 80 and over, which are the official estimates. We reconstruct our own population estimates at ages 80 and over.

So we use different methods below age 80 and above age 80, and above age 80 we use three different sets of methods, depending on the age reached by each cohort at the point of the most recent data available. So for some of these cohorts, we can assume that they're extinct because they have reached an age that we've seen in the past, for that particular population, extinction. So for this cohort, we use the extinct cohort method.

For cohorts which have reached the age of 90—so let's say that we're working on Canada, and the latest year available is 2011—for all the cohorts which have reached the age of 90 in 2011

but which cannot yet be considered extinct, we use something that we call the survivor ratio method. And I'll try to describe that really quickly, although I don't have a lot of time. And for all the cohorts which have not yet reached the age of 90 by the last point of data available, we us thee intercensal survival method in the same way as for the population below the age of 80.

So this is a graphic that summarizes what I just said. So for all the cohorts and ages in this A shape, we use the intercensal survival method. For all the cohorts here in this B shape, we use the extinct cohort method. And for all the cohorts in the C part, we use the nearly extinct cohort method, which is the survival ratio method.

So below age 80, we first redistribute population of unknown age. We do that throughout the whole age range, and then if we don't have January 1 population estimates—like in the U.S., we only have July 1 estimates—we use a linear method to create the January 1 estimates that we need. And then, typically, we would use nationally produced population estimates, single-calendar-year estimates. They're not always available, and even when they're available, they're not always reliable, although they are not always updated with the most recent data available. So I'm not going to go into details here.

So the intercensal survival method we use depends on the cohorts. So for preexisting cohorts and for new cohorts, we use slightly different methods, though the idea behind [them] is the same. So let's say we have two censuses, one at time t and one at time $t + 5$. We can follow a cohort—and at this point, we have reconstructed death counts by Lexis triangles, so we have deaths at a very detailed level—so we just age the cohort by removing the deaths in the triangle. And we have here a population age x at time t, and then by removing the deaths, we move it to January 1 of the following year, and we keep going on until we bridge with that second point of data, that second census. Of course, there are discrepancies because of errors both in terms of coverage and age reporting, but also because of migration. And at this point, if we don't have—which is mostly the case—detailed information about migration by year and ideally by age, we just assume that the difference between the reconstructed and the observed population size here for this age group and this cohort represents these migration and error quantities that we just redistribute uniformly throughout the cohort. And we do the same, starting from birth, for cohorts which are born during

the intercensal interval.

So again, we do that when we don't have official estimates, but we also do that sometimes when we have official estimates but they have not been updated. So typically, a country would produce postcensal estimates for the years since the last census, but in some cases, and most of the times, the countries adjust these estimates once a new census becomes available. There is retrospective adjustment for the intercensal period, but not all countries do so, and here are a few examples of countries which do not adjust when new data become available. So we produce our own intercensal estimates sometimes with help from statistical offices. So if we don't adjust, you see here discrepancies between two successive periods of estimates.

For a population at ages 80-plus, the process is much more complex. Again, we estimate—whether we use the extinct cohort method or the survival ratio method—we estimate, basically, population size by accumulating the deaths back, and this is pretty straightforward for extinct cohorts. I think I may have a graph here. So what we do is that in extinct cohorts, we know that by summing all of the deaths in a cohort, we can reconstruct the size of the cohort at various ages. So we reaggregate backward the deaths to estimate population at January 1 in that cohort at every age and every calendar year. So this is very straightforward, but for cohorts which are not extinct, we use a different method called the survival ratio method, which relies on the ratio of the cohort size at two successive ages, and we use information from previous cohorts for the same population, with a coefficient adjusting for change in mortality, and this is this *C*-coefficient here.

So I'm not expecting you to understand all of the details of the methods I'm presenting. I'm just trying to show you how complex the whole process is and how many decisions, in terms of the methods we have to take throughout that process, to really emphasize the fact that other decisions could have been made and would have produced different results. And so this is just a warning to be careful when you interpret the data.

So the next steps are much more straightforward. So from this, we now have population estimates and population by Lexis triangles, since we have the deaths by Lexis triangles. And we can use this method to reallocate the population throughout the cohorts.

So this was the example of what happens when we assume uniform distribution of births

throughout the calendar years. So this example is for France, and you see here a diagonal. So what's represented here are death rates for each single year of age and single calendar year in comparison with the death rates for the same age, the same single year of age, for the year before. And so the rate diagonal here indicates that for this particular cohort, the death rates are much higher than for the year before, and for this cohort here, this blue diagonal, the death rates are much lower than for the year before. And what happened here is that our hypothesis—our assumption of a uniform distribution of births—is severely violated, and you see here why. This is the distribution of births by month in France during the period around World War I, so you can see that in 1915 and again in 1919, there were huge fluctuations throughout the year. The number of births during World War I declined by about half in France compared to the month before and after the war, and there is, of course, a lag of nine months, which is why the cohort affected are those for 1915 and 1919, and not for 1914 and 1918, when the war actually took place.

So we've introduced this new method in version 6 of our protocol, which is going to be published very soon. We're still checking a number of things, and this method takes unequal distribution of births throughout the years in a way to adjust the measure of exposure.

Okay, so I don't have time to get into details, but if you have specific questions, I can show you a few graphs.

So we only use summarized information about the distribution of births—basically, the mean time during the year when the births occurred. So if this mean is 0.5, it means that the births are uniformly distributed and can vary between 0 and 1, and then we use the variance of the distribution. And we have particular equations which rely on the same principle for both the period and the cohort calculations. So, again, this is to adjust the exposure.

Once we have the death counts by Lexis triangles and the exposures by Lexis triangles, we take the ratio of one to the other to compute death rates. But we do one more thing, which is to smooth the death rates at high ages to deal with large fluctuations in small numbers. So this is an example of what happens at very high ages because of the very small numbers at these high ages and we're interested in the underlying mortality curves. We don't want large swings from year to year in our estimates due to, again, very small numbers. So we smooth, using a method developed

by Kannisto, which is basically a fit of the logistic function with an asymptote at 1 to make sure everyone ends up dying at the end. So this is a graphical representation. So this is the Gompertz that we would have if we only applied the Gompertz formula, and here is the asymptote at 1. And we validated this method using a number of examples; this is, for instance, Sweden in 2000. You can see that the fit is really good and they're only at these points at very high ages, which are outliers.

So this is just an opportunity to issue a very strong warning. There have been people who used the HMD to study mortality at very high ages, above 100 or 105, and this is absolutely not a good idea. First, because of the redistribution we have to do for most time periods in most of the countries, we do not have death counts to the highest possible age. We have death at high age combined into an open age interval, and we reconstruct what we think is the most plausible distribution of deaths above that age, and then we additionally have these smoothing methods at very high ages, which make the estimates what they actually are—estimates of mortality. And so, again, do not use the HMD to study mortality at very high ages. There are other databases out there—in particular, at the Max Planck—which are much more reliable in this respect.

And then finally from the death rates, so smoothed for very high age, we construct the complete life tables by single years of age, and from the single-year-of-age life tables, we construct the abridged life tables by five-year and 10-year age groups to ensure consistency. And we carry this process separately for men and women, and at the very end, we combine the information with, of course, weights to account for the differential distribution of men and women in the population. So we use the separate sex-specific life tables to construct the combined-sexes life tables.

So I'm going to pass on that.

And so, next, again, we update the documentation. We check everything through this extensive checklist, and we have a number of graphs that are automatically produced to help country specialists evaluate the quality of the output. Everything is checked through the two teams, and then finally, we publish the final results.

So I'm sorry I've been a little bit long. This work has only been possible because, again, of the help of a large number of individuals within and outside the HMD and the support of many

different institutions, including the Society of Actuaries and other insurance and reinsurance companies and professional organizations. So again, we are very grateful for your support and for your interest in the HMD. I'm going to stop here.

DALE HALL: We have about 10 minutes left to answer any questions, and we'll keep the presentation up in case we need to flip back to a certain graph or a formula. Please use the microphones in the center of the room, and we'll start here.

MAGALI BARBIERI: Just one. I'm sorry to cut you, but we're always open to answer questions, so don't hesitate in the future if something comes up, and either following this presentation or in your own work, to contact us directly. There are a few email addresses on the website, and we always answer our users' questions, so feel free to continue this conversation through email. Yes, I'm sorry.

LARRY PINZER: Larry Pinzer, Aon Hewitt. Magali, thank you very much.

MAGALI BARBIERI: Sure.

LARRY PINZER: Steve Goss yesterday alluded to the fact that mortality rates that HMD is showing for the USA are noticeably different at ages 65 and above what SSA is showing. Can you comment on that a little bit?

MAGALI BARBIERI: Absolutely. So we're actually working on that together with the Social Security Administration. So there are three sets of national life tables that are extensively used in the U.S. There's the Social Security Administration Trustees Report life tables. There are the NCHS [National Center for Health Statistics] life tables. The NCHS, for those who are not from the U.S., is the organization in the U.S. that collects, processes and disseminates mortality data for the whole country. And there are, of course, the HMD life tables. And there are discrepancies between the three, with NCHS falling in between SSA and HMD.

So there are two reasons why there are differences, and indeed, there are differences above age 65 that are not insignificant, especially when you do projections, because, of course, the differences accumulate over time and translate into wide discrepancies. So the differences are due first to the data. The data used by the Social Security Administration for ages above 65—. So below age 65, they base their life tables on the same data as the HMD, which are the NCHS data,

and differences are negligible. Above age 65, that's where the differences become increasingly important as we go up in age. The data used by the SSA are not vital-statistics data, as we do. They are data from the CMS, which is the Centers for Medicare and Medicaid—basically, Social Security Administration data—which have two big advantages. The first one is that there is consistency between the numerator and the denominator, because both the deaths and the exposures are computed from that data.

And, second, there has been some work done by a member of the Social Security Administration, Bert Kestenbaum, that showed that the quality of the age reporting in Social Security Administration data is better than in vital statistics, though this quality has recently very much improved in vital statistics, and age misreporting is mainly now an issue limited to certain sub-tracts of the population, mostly areas with high proportions of low education and with high immigrant populations, because these migrants come from countries where vital-statistics systems are often not working very well, and so the age reported is not always very accurate. But it's also true for older cohorts in the U.S. The U.S. Vital Statistics System achieved full coverage in 1933, so people born before that time in particular states don't always have good age registration or birth registration information.

We're working with SSA right now to try to allocate—. So the second source of differences is the methods that they are using; they are using very different methods than we do. They also do some smoothing at higher ages above age 95, like we do, but in a different way. So there are differences in data and differences in methods, and we are trying right now to work on trying to allocate the difference between our two sets of life tables to either the data or the methods, because the data of the Social Security Administration or CMS are much more reliable, but it's not complete coverage. They don't cover the whole population, and so there is also a selection effect that might bias the results if one is interested in really the national population. So we're trying to reconcile. And the NCHS is using a mix of methods that sort of falls in between. Some of the methods are closer to SSA; some of the methods are closer to HMD. And they use the same data as SSA, but their life tables, especially for the period since 2007–2008, their life tables are much closer to HMD than to SSA, even though they use the SSA data at higher ages, so not all of the

difference can be attributed to the data. Some of the difference is attributable to the method.

And again, I want to reemphasize here the fact that we know that there might be much better methods for a particular period for a particular country, but because we want to achieve comparability, it's very important that we can stand behind methods that can be applied across the board to all of the periods and all of the countries in that particular data situation.

DALE HALL: Let me go to the second microphone. John.

JOHN ROBINSON: Thank you, Dale. John Robinson, Minnesota Department of Commerce. Just two general questions. First of all, thank you very much for the presentation, and it's quite amazing how something that seems simple to begin with—let's count deaths and divide by exposure—gets complicated real fast.

Just two general questions. Each year, you get the data, and you create a mortality table. Do you ever go back and recompute a table based on emerging information about past times?

MAGALI BARBIERI: We absolutely go back. So there are different situations in which we recalculate our life tables for the past, using new data. So one situation is when a country issues one more year of data. So, for instance, for the U.S., we now have mortality data for 2015. We're waiting for the population data. We need January 1 population for 2016. That's not been published yet but should be published this month, from what I'm told. As soon as this is available, we will update the U.S. for 2015.

By doing so without changing input data for any other year, but by doing so, because of the extinct cohort method and the survival ratio method we're using, this is going to modify—adding a new year of death information is going to modify the population estimates or the exposure estimates we have for the years for which these cohorts were not extinct. And so that's going to modify the life table estimates for 10 or 15 years' worth of data.

There is also a situation where, again, when there is a new source of population data—a new census, typically—statistics offices substitute intercensal in most countries—not all, as I showed you—they substitute intercensal estimates to their previously postcensal estimates, and so we use this new data. And then there are instances where countries carry out a complete revision of previous estimates, and that happens on a regular basis. So I think the most recent case was one

of the Scandinavian countries—I believe it might have been Sweden—where they reconstructed more carefully population estimates for historical periods. And, again, because we have the special relationship with statistical offices, they often spontaneously come to us and tell us, "Look, we have this new set of data; could you use these instead?"

JOHN ROBINSON: And so when you do sort of restate a previous table, do you take out the old table, or do you keep both?

MAGALI BARBIERI: So we do keep the old tables in the database, but they are not accessible to the public, because that would be really confusing.

JOHN ROBINSON: I see. Okay.

MAGALI BARBIERI: But we do document each update, or unless they are pretty straightforward, but we document in the country-specific documentation. So we have some general documentation, like these methods particularly that I showed you, that's applicable to countries. We also have very detailed country-specific documentation, and in the country-specific documentation, you will find information about what has changed. So, for instance, for this example I was giving of Sweden, that's very specifically mentioned in the documentation that there has been a change in the population data from the 19th century, and I think it was up to the first half of the 20th century, for this particular country.

It's happened very rarely, but it's happened that people have contacted us and said, "Look, I used your data to publish that paper three years ago. The data is now different; I have to make some revision; I don't want to redo all of the analyses. Could you send me this set of original data that I used three years ago?" And we were able to do that, because we have a version-control system. We keep all of the estimates we have constructed over the years.

DALE HALL: We may have time for one final quick question here at the front, if that's okay.

NATALIA GAVRILOVA: Natalia Gavrilova, NORC/Chicago. You showed that you smooth the death rates by logistic formula after age 80 for period death rates. And do you do the same for cohort death rates, smoothing by logistic formula? I'm asking this because there was a study, that used cohort death rates and showed that, again, the Kannisto model or logistic formula is the best..

MAGALI BARBIERI: So let me answer with two different points. The first point is we do not

smooth the death rates for the cohort life tables. We also make available to the users the period death rates unsmoothed, which is the big $M(x)$ series, but the central death rates in the life tables, the little $m(x)$, for the period life tables only are smoothed, and the smoothing is not done systematically at age 80. The smoothing—so I didn't have time to mention that, but it's really a fit rather than a smooth, but the logistic function that we used for the smoothing at higher age is run over the whole age range, but the smoothed death rates are substituted to the observed death rates at ages above 80, depending on the number of population remaining. So we pick the age at which there are at most 100 males or 100 females, unless this number is reached after age 95, in which case we use 95 as the minimum age where we substitute the smoothed rates to the observed rates.

NATALIA GAVRILOVA: Is it for all countries subset?

DALE HALL: I'm sorry, I'm going to need to end this here. We're at the top of the hour. You can come up afterward. Thanks to Magali for joining us here, and we'll move to our next sessions. Thanks, everyone.

MAGALI BARBIERI: Thank you.