MILLIMAN WHITE PAPER

# The Withdrawal Delay Cohort Method under VM-21/AG-43: The case for random sampling

Benjamin Buttin, ASA, MAAA
Matthias Kullowatz, ASA, MAAA
Zi Xiang Low, FSA, FIA, MAAA
Zohair Motiwalla, FSA, MAAA

Milliman

## Background

In early December 2017, the National Association of Insurance Commissioners (NAIC) released proposed revisions to the existing U.S. variable annuity (VA) statutory framework. These revisions were promulgated as redline updates to the existing Actuarial Guideline 43 (AG-43) and Risk Based Capital (RBC) C3 Phase II (C3P2) instructions, paving the way for VM-21 of the Statutory Valuation Manual (VM), Requirements for Principle-Based Reserves for Variable Annuities. After an exposure period in early 2018 to allow for comments from industry participants, regulators, and interested parties, the Variable Annuity Issues (E) Working Group (VAIWG) of the NAIC adopted almost all of the recommended changes outlined in the redline instructions.

While these revisions have been broadly agreed upon by the NAIC, a final set of regulatory instructions for VM-21 is still pending, with the responsibility assigned to the VM-21 Report Drafting Group. New updated redline instructions are exposed publicly on a piecemeal basis, inviting comments and feedback from practitioners and interested parties.[1] The working expectation is that the final version of VM-21 will be formally adopted at the NAIC Summer meeting in August 2019, for a January 1, 2020, effective date. Under the new VM-21 framework, the Aggregate Reserve is now the sum of the Conditional Tail Expectation 70 Amount (CTE Amount) and the Additional Standard Projection Amount, where the latter term is determined using the Standard Projection.

## The Standard Projection

The VM-21 Standard Projection is essentially a complete overhaul of the existing AG-43 Standard Scenario framework. It can be calculated using either the Company-Specific Market Path (CSMP) method or the Conditional Tail Expectation with Prescribed Assumptions (CTEPA) method. The CSMP method uses (at least) 40 prescribed economic scenarios while the CTEPA method uses the same (number of) economic scenarios as the CTE Amount calculation (common practice is to use at least 1,000 scenarios).

Companies might find the CTEPA method desirable because it uses the same real-world economic scenario set as under the CTE Amount (although of course with prescribed assumptions) and so provides an intuitive and commensurate comparison. In particular, we note that using the CSMP method creates a dependency between the CTE Amount and the Standard Projection because the former must be determined first before the latter can be calculated.

## Withdrawal Delay Cohort Method

One of the more challenging (and key) components of the Standard Projection is the Withdrawal Delay Cohort Method (WDCM), which is a prescribed approach for determining the timing of policyholder election for policies with either hybrid guaranteed minimum income benefits (Hybrid GMIBs)[2] or guaranteed minimum withdrawal benefits (GMWBs). The WDCM applies in both the CSMP and CTEPA methods. To be in scope for the WDCM, policies must be either nonconforming (meaning they have taken a withdrawal in the policy year occurring coincident with the valuation date, and this withdrawal was in excess of the GMWB's guaranteed annual withdrawal amount or the GMIB's dollar-for-dollar maximum withdrawal amount) or non-withdrawers (meaning that they have not started taking withdrawals).

Under the existing AG-43 framework, the Standard Scenario assumes that the exercise of any living benefits such as GMIBs or GMWBs occurs at the earliest available opportunity that is consistent with contractual provisions.

In contrast, the WDCM under VM-21 defines a prescriptive process for determining a distribution of possible election cohorts for each policy in scope, each with its own weight. The cohorts simulate each potential age of starting systematic withdrawals. In order to determine the election distribution, the Guaranteed Actuarial Present Value (GAPV) concept, as prescribed under VM-21, is

[1] This article has been developed using the updated VM-21 redline that was exposed in early March 2019. The reader is cautioned that, to the extent that the final version of the instructions is different from this redline, certain content in this article may need to be revised.

[2] A Hybrid GMIB policy is a policy with both guaranteed growth (such as with a rollup or doubler) and dollar-for-dollar partial withdrawal reductions in the GMIB benefit base.

used to calculate the prospective withdrawal value of the rider to the policyholder at each potential individual withdrawal age.

The main steps in the WDCM are outlined below:

- For each potential initial withdrawal age (from issue to attained age 120, subject to contractual provisions), compute the GAPV assuming the policyholder elects to take withdrawals at that age. This will produce a set of GAPVs.
- Apply certain prescribed transformations and normalizations to this set of GAPVs to develop a from-issue cumulative distribution function (CDF), reflecting shocks as necessary.[3] This CDF defines a specific weight for the withdrawal cohort corresponding to each initial withdrawal age from issue.
- A "never withdraw" cohort is also defined, whose weight varies by rider type and tax status.
- Given a valuation date, any withdrawal cohorts corresponding to initial withdrawal ages occurring prior to that date are discarded and the remaining weights are rescaled to produce a revised CDF (call this the "rescaled CDF").

The key drivers in this process are those that underlie the GAPV calculation, namely the rider benefit base mechanics, the payout rate for the GMWBs and/or Hybrid GMIBs under consideration, the prescribed Standard Projection mortality, and the discount rate (3%). The most recent redline instructions stipulate that the CDF is calculated once for a set of policies with the same combination of issue age, rider type, and tax status. For the purposes of this article, we refer to this combination as the "WDCM cell key." In practice, there may be legitimate reasons to expand the WDCM cell key definition. For example, gender is a key item that should also be considered (because mortality rates will vary by gender). Moreover, the payout rate may vary by joint life status or rider generation.

Theoretically, policies with the same WDCM cell key should produce the same from-issue CDF even if their benefit bases on the valuation date are different, because the associated GAPVs should simply scale and the weights would renormalize to the same values. One could even calculate the CDF using an arbitrary (but non-zero) benefit base amount. Accordingly, for existing policies the calculation of the from-issue CDF is intended to be a one-time process—once calculated for a given WDCM cell key, the weights are fixed and do not need to be recomputed in the future.[4] The practitioner need only compute new weights for new business issued that have different WDCM cell key combinations.

While the WDCM process is theoretically very appealing, in practice the run-time associated with splitting the in-force file into many cohorts (some of which may be assigned very small weights) can be very challenging, particularly under the CTEPA method. The full WDCM cohort file record count is likely to be many times greater than that of the original in-force file.

The redline instructions provide some allowance for discarding additional cohorts to mitigate the computational burden, so long as this decision has been disclosed. The specific language indicates that individual withdrawal age cohorts may be discarded or a small number of withdrawal cohorts may be assigned to each contract via random sampling.

Discarding cohorts to relieve the computation burden without loss of accuracy (relative to results produced using the full WDCM cohort approach) requires practitioners to engage in some analysis and testing, ideally before VM-21 becomes effective. There are a number of approaches that companies might take. For example, companies could (a) specify a maximum number of cohorts (and map cohorts with the smallest weights to those with the largest weights), (b) collapse all the cohorts for an in-force policy down to a single cohort by using the weighted average deferral time across all cohorts for that policy, or (c) similar to (b) but using the median deferral time. Other reasonable approaches may prove to be suitable as well.

As noted in the redline instructions, one possible route practitioners can take is to use a random draw to collapse all cohorts to a single cohort for each in-force policy. The process would involve using a robust random number generator to produce a random draw on the interval 0 to 1 for each in-force policy. This value would be compared to the rescaled CDF produced by the WDCM process, thereby randomly selecting a future election time and modeling each in-force policy using a single cohort with that particular election time. The advantage to this approach is that the in-force file record count for the randomized run is the same as the pre-WDCM version (i.e., the original in-force file). For proof of principle, the practitioner should verify that the results produced using both the random sampling approach and the full WDCM cohort approach are not only similar, but that repeated random trials produce stable results. This test should be performed at the onset of adopting the random sampling approach, and may also need to be carried out at future intervals (such as to support disclosure of the approach in the year-end Actuarial Memorandum).[5] It should be noted that a number of companies already employ random sampling methods in their CTE Amount calculations.

[3] For applicable policies, these prescribed shocks correspond to the end of the rollup period and/or required minimum distributions after age 70 for qualified plans.

[4] Other than for the rescaling as the valuation date changes. Also, if there is a model correction/refinement that impacts the key drivers outlined above, then the CDFs need to be recalculated. An earlier version of the redline-specified mortality improvement to the valuation date for the GAPV, which would have resulted in varying CDFs over time (if they have been recomputed at future valuation dates). A recent change to the redline to reflect improvement for the GAPV to December 31, 2017 (rather than the valuation date) avoids this situation.

[5] Another test that can be performed to make sure that the functionality is correct would be to run both approaches and force all policies/cohorts to use a single election time. One would expect both approaches to produce near-identical results.

## Statistical theory behind random sampling

In defense of the random sampling approach outlined above (in which a single delay cohort is randomly selected for each policy) we argue that the Greatest Present Value of Accumulated Deficiencies (GPVAD) calculated by randomly sampling the election time for each in-force policy will converge to the true GPVAD within an economic scenario for large in-force sizes, where the true GPVAD is that which would be calculated by using the full WDCM cohort in-force file. We start by showing convergence of the policy-level accumulated product cash flows, and we expand that to the convergence of the GPVAD.

Probability theory suggests that when you sample values from a population, the ratio between the sample standard deviation and the sample sum shrinks as the sample size increases. The sample standard deviation here can be thought of as an "error," the discrepancy between our GPVAD estimate and the true GPVAD. As such, even though larger in-force sizes will generally lead to larger errors, the errors will actually become smaller as a proportion of total GPVAD.

This theory extends naturally to WDCM cohort sampling—which is effectively a form of stratified sampling—where exactly one outcome is randomly selected for each policy. We first conceptualize the effect using the policy-level accumulated product cash flows. Each policy has a theoretical variance of possible accumulated product cash flow values, based on the randomness of which WDCM cohort is sampled. Because WDCM delay cohorts are sampled independently for each policy, the variance of the sum is equal to the sum of the variances, shown mathematically below:

$$Var\left(\sum_{1}^{n} X_i\right) = \sum_{1}^{n} Var(X_i)$$

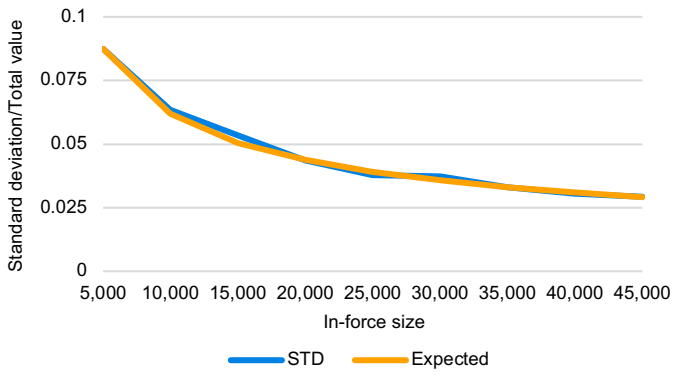
$$\text{where } X_i = \text{sampled cash flow value for } i^{th} \text{ policy}$$
$$\text{and } n = \text{in} - \text{force size}$$

As such, the variance of the sum increases linearly with the in-force size, implying that the standard deviation of the sum increases at a rate proportional to the square root of the in-force size. In other words, the sum is growing at a linear rate but the standard deviation, or "error," is growing at the rate of the square root, which is much slower.

In order to illustrate this relationship, we started with nine sets of in-force files that contained samples of between 5,000 and 45,000 policies. Each of these in-force files contained policies that were cohorted under the prescribed full WDCM approach with accumulated product cash flow results pre-calculated for each cohort. For each of these in-force files, we randomly sampled distinct sets of cohorts 1,000 times to generate a distribution of potential total accumulated product cash flows.

In Figure 1, the blue line represents the ratio of the standard deviation of the random samples to the total accumulated product cash flows for each in-force file size, while the orange line represents the ratio that we would expect to see if the square root principle held. The graph shown in Figure 1 explains the phenomenon nearly perfectly. In other words, the sample error—as measured by the sample standard deviation—will shrink at a rate proportional to the square root of the in-force size.

**FIGURE 1: RATIO OF STANDARD DEVIATION TO TOTAL ACCUMULATED PRODUCT CASH FLOWS BY IN-FORCE SIZE**



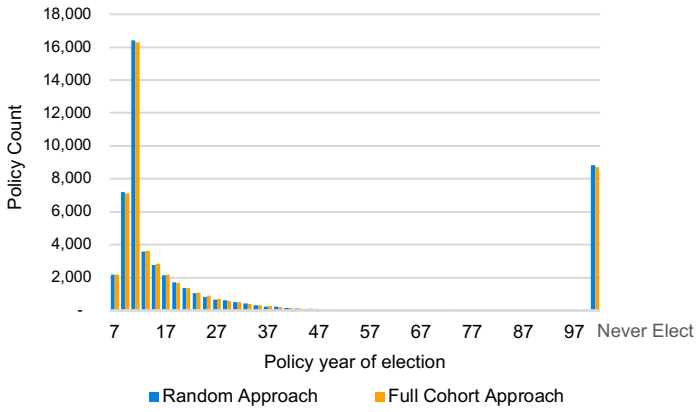

While the probability theory discussed above explains the variation for sums of policy-level cash flows quite well, it does not cover how convergence of a policy-level cash flow implies convergence of the GPVAD. Intuitively, the calculation of GPVAD implies additional aggregation, both within and across time steps, and aggregation generally leads to lower variances. For example, this concept of aggregation is used to diversify portfolios and reduce risk. The case study in the next section supports the hypothesis that the relative variation in GPVAD across random samples is lower than the relative variation of policy-level cash flows, suggesting that using policy-level cash flows is a conservative approach to determining whether an in-force size is large enough.

## WDCM case study

For our case study we implemented both the full WDCM cohort approach and a random sampling approach for a block of roughly 52,000 VA policies with guaranteed lifetime withdrawal benefits (GLWBs) and $6.5 billion in account value in-force. These GLWB policies have an annual account value ratchet, a 5% benefit base rollup for the first 10 policy years, and a maximum annual withdrawal amount ranging from 3% to 6% by attained age. On implementation of the full WDCM cohort approach, the in-force size increased by an order of tenfold to just under 590,000 cohort records.

We performed a random sampling of the withdrawal election times for the block of GLWB policies and compared the resulting distribution against that produced by the full WDCM cohort approach in Figure 2 below. For the random sampling approach, the y-axis represents the total policy count for each year of election. For the full cohort approach, the y-axis represents the sum of the probability weights across all cohorts assigned for each election time. Because the sum of probability weights is equivalent to the expected value of each election time, this gives us a metric against which we can compare the sample counts. It is immediately evident from Figure 2 that the shape of the distribution is very similar between the two approaches (as expected, due to the law of large numbers and the probability theory discussed earlier).

**FIGURE 2: DISTRIBUTION OF WITHDRAWAL ELECTION TIMES (POLICY YEAR)**




Next, we projected the aggregate cash flows for both approaches using the prescribed VM-21 assumptions for the Standard Projection under a single adverse tail-end economic scenario. We plotted the movement of aggregate free surplus across all projection periods under both approaches and noted the near exact match (Figure 3). In addition to that, the table in Figure 4 shows that the dollar difference in free surplus between both approaches is minimal at various projection quarters.

**FIGURE 3: FREE SURPLUS (IN $) ACROSS PROJECTION QUARTERS FOR SINGLE ADVERSE TAIL-END SCENARIO**

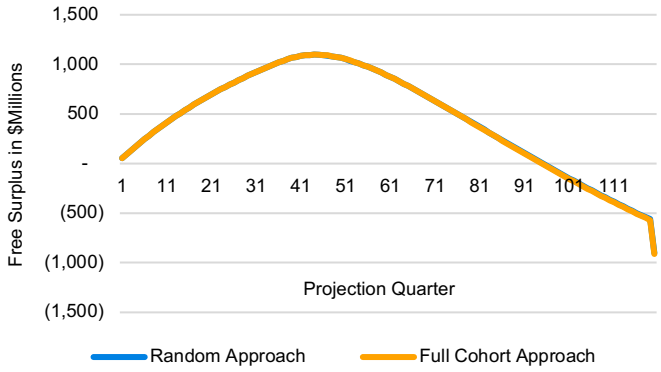



**FIGURE 4: FREE SURPLUS (IN $) AT VARIOUS PROJECTION QUARTERS FOR A SINGLE ADVERSE TAIL-END ECONOMIC SCENARIO**

| PROJECTION QUARTER | FREE SURPLUS | | |
|---|---|---|---|
| | FULL COHORT APPROACH | RANDOM APPROACH | % DIFFERENCE |
| 1 | 52,804,571 | 52,804,774 | 0.0% |
| 25 | 794,526,441 | 794,578,862 | 0.0% |
| 50 | 1,064,869,166 | 1,064,191,842 | -0.1% |
| 75 | 525,080,878 | 527,427,921 | 0.4% |
| 100 | (130,906,609) | (124,967,114) | -4.5% |
| 120 | (911,403,069) | (901,397,103) | -1.1% |

From the table of results in Figure 5, the GPVAD results between the full WDCM cohort approach and the random sampling approach are very close for three tail-end economic scenarios. In order to validate the stability of the random sampling approach, we performed additional random samplings (with a different random seed each time) and recalculated the GPVAD result. As is evident from the table in Figure 5, the random sampling approach produces very stable results across all three scenarios.

**FIGURE 5: GPVAD RESULTS (IN $)**

| | GPVAD | | |
|---|---|---|---|
| | 85TH PERCENTILE | 95TH PERCENTILE | WORST SCENARIO |
| Full Cohort Approach | (513,763,374) | (293,441,348) | 362,313,321 |
| Random Approach Run 1 | (508,157,194) | (286,549,708) | 373,580,935 |
| Random Approach Run 2 | (515,788,554) | (295,226,691) | 360,822,887 |
| Random Approach Run 3 | (512,286,245) | (291,652,102) | 365,459,258 |
| Random Approach Run 4 | (513,587,416) | (292,779,601) | 364,035,113 |
| Random Approach Run 5 | (514,675,793) | (294,887,823) | 358,851,912 |
| Mean for Random Approach | (512,899,041) | (292,219,185) | 364,550,021 |
| Ratio of Standard Deviation over Mean for Random Approach | -1% | -1% | 2% |

# Other considerations

Companies can calibrate their own policy-level cash flow variance and corresponding GPVAD variance for small sample sizes of their in-force blocks, and then extrapolate out to see whether randomly sampling cohorts on the full in-force is expected to achieve a tolerable error. The following methodology provides an example of how a company could determine the required in-force size to achieve its desired error rate, where error is defined as the ratio of the standard deviation of GPVAD values across random cohort samples to the true GPVAD value of the block:

1. Randomly sample a manageable subset of policies of size N.
2. Randomly sample one delay cohort for each of these N policies. Calculate the sample GPVAD value, then repeat random draws many times, say 100, to produce 100 random GPVAD values.
3. Calculate the means and standard deviations for each set of 100 values, $\bar{G}$ and $S_G$.
4. Assume that the ratio of relative GPVAD error to the relative error of total policy-level cash flows remains constant, and then we can use the theory presented in the Statistical Theory section above: as the in-force size multiplies by a factor of F, that the relative error will shrink by a factor of $\sqrt{F}$.
5. Let X be the target sample size to achieve a desired GPVAD relative error, and let that desired GPVAD relative error tolerance be T. Solve the following equation for X:

$$\frac{S_G}{\bar{G}} \cdot \sqrt{\frac{N}{X}} = T$$

As an example, if a company were to randomly sample 5,000 policies from its in-force, and then generate a relative GPVAD sampling error ($\frac{S_G}{\bar{G}}$) of 5% across 100 distinct cohort samples, it would require an in-force size of 125,000 policies to achieve an approximate GPVAD relative error of 1%. These numbers are for illustration only.

Note that as more GPVAD values are sampled in Step 2, the sample mean $\bar{G}$ becomes a closer and closer estimate to the true GPVAD value for the selected subset of policies. By taking 100 samples, probability theory tells us that the standard error of the sample mean will be approximately equal to the standard deviation of the samples divided by 10, the square root of 100. This square root rule holds for all potential sample sizes. By taking 100 random samples to calibrate our first point, we guarantee that the standard error of the sample mean $\bar{G}$ is only about 10% of the standard deviation of GPVADs, assuring that our estimate of the relative GPVAD sampling error ($\frac{S_G}{\bar{G}}$) is a precise estimate. The practitioner can also choose to run all cohorts on this subset of policies to obtain a true value for GPVAD for comparative purposes.

There are some extreme cases where even fairly large in-force sizes cannot completely immunize the simple random sampling process from intolerable variation. In cases where a few policies contribute disproportionately to the metric of interest, the selection of a WDCM cohort for those few policies will also disproportionately affect the variance. If a company is unable to smooth out the variance due to such skewed distributions, it can choose to model all WDCM cohorts for its riskier business blocks, while using the random approach to sample WDCM cohorts for the other business blocks. This approach could be used on a smaller entity that exhibits such skewness in cases where reserves are reported at the level of the legal entity and/or sub-group.

Aside from the number of policies, the speed of convergence will depend on how materially different the probability distributions are between distinct WDCM cell keys. For example, if the contribution to the GPVAD result at each initial withdrawal age were not materially different among a group of WDCM cell keys within an in-force file, we would expect the randomly sampled results to converge much faster than a similarly sized in-force with vastly different WDCM cell key distributions. However, in the latter case companies can attempt to identify underlying characteristics of the WDCM cell key that produce high variances and adjust their sampling methods to achieve better convergence (for example, leveraging stratified sampling by selecting more than one cohort per policy).

Lastly, when considering the random sampling approach companies should set seeds for each random draw for the sake of reproducibility. To preserve independence between unique policyholder decisions, it is important that these seeds are unique to each policyholder. Additionally, companies may wish to set distinct seeds across economic scenarios and perhaps even across valuation dates. By selecting unique seeds across policyholder, economic scenario (and, potentially, valuation date), practitioners can reduce overall bias from the random sampling method within valuation dates and across them. Of course, it is important to first implement this modeling approach and analyze these considerations in a test bed environment before moving them to production.

## Conclusion

In recognition of the potential run-time challenges posed by the Withdrawal Delay Cohort Method for variable annuity statutory valuation requirements under the VM-21 Standard Projection, we expect that companies will be looking to incorporate innovative solutions to manage the computational burden. Random sampling offers one such solution, and one that is allowed within the proposed framework.



**CONTACT**

Benjamin Buttin
benjamin.buttin@milliman.com

Matthias Kullowatz
matthias.kullowatz@milliman.com

Zi Xiang Low
zixiang.low@milliman.com

Zohair Motiwalla
zohair.motiwalla@milliman.com